

# Creació d'un motor de TAE especialitzat en farmàcia i medicina per a la combinació romanés–castellà

## Creating an SMT engine for pharmaceutical and medical texts in the Romanian–Spanish language pair

Adrià Martín-Mor\*

Universitat Autònoma de Barcelona  
adria.martin@uab.cat

Víctor Peña-Irles

Universitat Autònoma de Barcelona  
victorpenairles@gmail.com

### Resum

Aquest article<sup>1</sup> descriu el procés de creació d'un motor de traducció automàtica estadística especialitzat en medicina per a la combinació lingüística romanés–castellà a partir de corpus lliures disponibles a internet. S'utilitza la plataforma MTradumàtica, creada en el marc d'un projecte de recerca del grup Tradumàtica per a fomentar l'ús de la TA entre els traductors. L'article es pot interpretar com una mostra que aquest propòsit s'ha assolit en el cas d'ús que presentem, la qual cosa suggereix que el perfil dels traductors és vàlid per dur a terme processos de personalització de TA.

### Paraules clau

traducció automàtica, traducció automàtica estadística, personalització de motors de traducció

### Abstract

This article<sup>2</sup> describes the process of creation of a statistical machine translation engine specialised in medicine for the Romanian–Spanish language pair. The engine was based on free corpora available in internet. The article describes the use of the platform MTradumàtica developed in the context of a research project by the Tradumàtica research group, aimed at promoting the use of MT among translators. The article can be interpreted as the evidence that the aim

\*ORCID: 0000-0003-0842-3190

<sup>1</sup>Els autors d'aquest article signen com a ciutadans de la República catalana proclamada pel govern legítim de Catalunya, en protesta per l'empresonament i exili d'activistes polítics i membres del govern i en solidaritat amb els ciutadans que van patir la repressió de l'Estat espanyol arran del referèndum d'autodeterminació de l'1 d'octubre del 2017.

<sup>2</sup>This article is signed, as citizens of the Catalan Republic proclaimed by the legitimate government of Catalonia, in protest against the imprisonment and exile of political activists and members of the Catalan government and in solidarity with all the citizens who suffered reprisals by the Spanish state following the Catalan self-determination referendum held on the 1st October 2017.

of promoting MT among translators has been attained in this particular case, and it suggests that the profile of the translators is valid to carry out processes of customisation of MT engines.

### Keywords

machine translation, statistical machine translation, statistical machine translation customisation

## 1 Introducció

En el marc de ProjecTA,<sup>3</sup> el grup de recerca Tradumàtica es va proposar analitzar l'estat de la traducció automàtica (TA) en el teixit empresarial de Catalunya i de l'Estat espanyol (Torres-Hostench et al., 2016). Els resultats de l'anàlisi van conduir a la creació d'una plataforma per a la personalització de motors de traducció automàtica estadística (TAE) per tal d'acostar la TA als professionals de la traducció. Aquest article descriu el procés de creació d'un motor de TA especialitzat en medicina per a la combinació lingüística romanés–castellà. L'article està dividit en 6 apartats. L'apartat 2 (Personalització de motors de TAE) descriu què és la personalització de motors i quines plataformes existeixen actualment per a aquestes tasques. L'apartat 3 (Recursos per a la creació del motor) avalua els recursos disponibles per a la creació del motor de TAE en la combinació lingüística esmentada, amb referències a altres traductors automàtics existents i una avaluació crítica dels punts forts i les febleses. Finalment, l'apartat 4 (Descripció del procés de creació) descriu els recursos utilitzats i la seua preparació, abans que els resultats i les conclusions (5 i 6 respectivament) tanquen l'article.

<sup>3</sup><http://www.projecta.tradumatica.net>. Referència FFI2013-46041-R, finançat pel Ministerio de Economía y Competitividad del Gobierno de España. Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad.



## 2 Personalització de motors de TAE

Actualment, diverses plataformes permeten la personalització de motors de TAE. En l'àmbit del programari privatiu, existeixen programes com ara KantanMT,<sup>4</sup> LetsMT,<sup>5</sup> Microsoft Translator Hub<sup>6</sup> o Slate Desktop (anteriorment, DoMosesYourself).<sup>7</sup> Moses, programari lliure amb llicència GNU Lesser General Public License,<sup>8</sup> és un dels programes més utilitzats per a la creació de motors de TAE. Segons LT-Innovate (2013, p. 71), Moses és “widely used within the industry to build customized MT engines” i, justament, es destaca que, com que es tracta d'una plataforma lliure, “people wishing to develop a custom engine can focus on obtaining the training corpora rather than writing their own statistical machine translation engine (a difficult task that is beyond the abilities of most developers).” Malgrat tot, tal com continua LT-Innovate (2013, p. 72), Moses és “difficult to administer”, començant pel fet que no té interfície gràfica d'usuari (GUI) i, per tant, requereix un cert coneixement de sistemes UNIX i del terminal, la qual cosa sol suposar una barrera d'entrada per a una gran part dels usuaris potencials. Probablement per aquest motiu, hi ha hagut en els últims anys intents de desenvolupar sistemes per a un públic menys expert en tecnologia. Per exemple, Machado & Fontes (2014) presenten un conjunt d'eines de programari lliure (desenvolupades “by a translator for translators”, p. 2) per a la creació de motors de TAE, com ara eines de conversió de formats o materials de suport. Més recentment, han aparegut sistemes basats en Moses amb interfície gràfica, com ara ModernMT,<sup>9</sup> Machine Translation Training Tool (MTTT)<sup>10</sup> o MTradumàtica, tots tres amb llicències lliures.

MTradumàtica<sup>11</sup>, actualment en versió experimental, és una plataforma web basada en Moses per a la creació de motors de TAE personalitzats (Martín-Mor, 2017). La llicència LGPL de Moses permet la modificació del codi font i la redistribució de programari, la qual cosa comporta que qualsevol usuari pot complementar o adaptar el programa original per als seus objectius, en el cas de ProjecTA, acostar la TA als traductors. A tal efecte, la plataforma desenvolupada es proposava:

1. Desenvolupar una interfície gràfica prenent en consideració una dimensió educativa envers l'usuari final.
2. Permetre l'ús via web, per tal d'evitar instal·lacions en local, la qual cosa converteix, *de facto*, el programa en multiplataforma.
3. Permetre la instal·lació en servidors propis, per tal d'assegurar una major confidencialitat en l'àmbit professional.

Des del punt de vista de la política de la recerca, el fet de contribuir al desenvolupament de programari lliure garanteix alhora que el producte de projectes d'investigació finançats amb fons públics esdevé també públic i disponible per a tota la societat.

Així, MTradumàtica segueix el següent esquema per a la creació de motors.

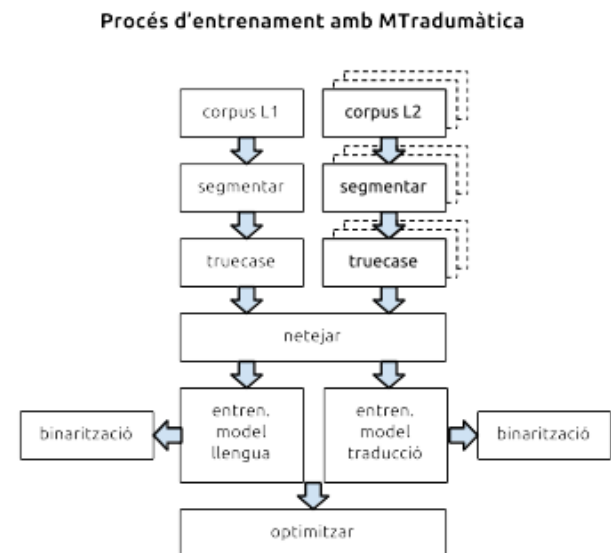


Figura 1: Esquema de processos en MTradumàtica.

A partir d'un corpus paral·lel bilingüe (per al model de traducció, en endavant, MT) i d'un o més corpus monolingües (per al model de llengua, en endavant, ML), MTradumàtica, com Moses (Koehn, 2016, p. 36), duu a terme els processos de segmentació (*tokenisation*), *truecasing* i neteja dels corpus. *Segmentar* vol dir separar amb espais les paraules dels signes de puntuació. En altres paraules, aïllar la puntuació permet incrementar les probabilitats d'obtenir coincidències amb els futurs textos que es traduiran automàticament. El procés de *truecasing*, en canvi, consisteix a determinar la caixa més probable de cada paraula, majúscules o minúscules. Tal com afirma Koehn al seu glossari de termes de Moses (Koehn, 2016, p. 361), “[t]his process typically leaves all words unchanged except for the

<sup>4</sup><https://www.kantanmt.com/>.

<sup>5</sup><https://www.letsmt.eu/>.

<sup>6</sup><https://hub.microsofttranslator.com/>.

<sup>7</sup><https://slate.rocks/>.

<sup>8</sup>Vegeu <https://www.gnu.org/copyleft/lesser.html>

<sup>9</sup><http://www.modernmt.eu>.

<sup>10</sup><https://github.com/roxana-lafuente/MTTT>.

<sup>11</sup><http://m.tradumatica.net>

first word in the sentence, which may be lowercased.” S’evita així que els vocabularis continguin entrades diferents per a la mateixa paraula en majúscules i en minúscules, i per tant les dades són menys esparses i es facilita l’entrenament. La neteja consisteix en la supressió de les frases llargues i mal alineades dels corpus amb l’objectiu de minimitzar els problemes en la fase d’entrenament.

Una vegada duts a terme aquests tres processos, el sistema processa les dades lingüístiques proporcionades en la fase d’entrenament, en la qual, a partir de l’anàlisi de coocurrències de paraules i segments en les dues llengües, s’infereixen de manera automàtica correspondències de traducció. El resultat de l’entrenament és el model de traducció, format per una taula de frases, un model de llengua i, ocasionalment, una taula de reordenament. Atés que la consulta de les taules pot ser lenta, els models es binaritzen per tal que es carreguen més ràpidament.

Finalment, l’optimització (o *tuning*) és un procés que determina automàticament els valors òptims d’una sèrie de paràmetres per tal que el motor generi “the best possible translations” (Koehn, 2016, p. 12). L’optimització consisteix en la traducció automàtica de milers de frases d’un subconjunt dels models (anomenat *development* o *tuning set*), la comparació amb les traduccions humanes de referència i l’ajustament automàtic dels valors de cada paràmetre per tal de millorar la qualitat del motor, mesurada mitjançant mètriques automàtiques com ara BLEU (Papineni et al., 2002). MTradumàtica es basa en els paràmetres per defecte de Moses per a fer l’optimització (no permet personalitzar-los ni tampoc té paràmetres diferents per a cada combinació lingüística). Un cop acabada l’optimització, el motor de TA estarà a punt. Actualment, MTradumàtica no permet dur a terme processos de postedició en la mateixa plataforma.

Pel fet que MTradumàtica ha estat dissenyat amb l’objectiu de facilitar l’acostament dels traductors a la TA, la interfície del programa conté referències als processos esmentats anteriorment. Tal com es pot observar a la Figura 2, malgrat que no és imprescindible tenir coneixements avançats sobre aquests processos per a la creació d’un motor de TAE amb MTradumàtica, l’eina també es proposa *formar* l’usuari en les nocions bàsiques de l’àmbit.

A tal efecte, la interfície inicial del programa presenta un procés lineal de sis passos (set, si es té en compte la funció *Inspect*, actualment en desenvolupament, v. més avall):

1. Càrrega de fitxers
2. Generació de monotextos
3. Generació de models de llengua
4. Generació de bitextos
5. Generació de traductors automàtics
6. Traducció

Els sis passos són visibles des de la pàgina inicial amb una breu explicació i indicacions addicionals. A més, al llarg de tot el procés, la barra superior mostra a l’usuari en quin pas es troba.

El procés comença amb la càrrega dels fitxers que posteriorment es faran servir per a la generació dels models de llengua i de traducció. De fet, la pàgina inicial, tal com es pot observar a la Figura 2, conté un enllaç al projecte Opus, el repositori de corpus lliures (Tiedemann, 2009). La pestanya *Files* mostra els textos carregats amb informació quantitativa (nombre de línies, paraules i caràcters) i la llengua del fitxer, detectada automàticament pel programa (l’usuari té la possibilitat de corregir la detecció automàtica en els casos en què falla). Davall dels textos carregats, hi ha un camp per a la càrrega de fitxers. En el moment de donar per tancat aquest article (desembre 2017), i tal com s’informa davall del camp esmentat, només es poden carregar fitxers de text amb una sola frase per línia.<sup>12</sup>

Al pas següent, *Monotexts*, l’usuari ha de generar monotextos amb l’objectiu de generar un model de llengua posteriorment, a la pestanya *LMs*. Es poden combinar diversos fitxers monolingües per tal d’obtenir un model de llengua més gran.

Un cop generat el model de llengua, la pestanya *Bitexts* permet —de manera paral·lela a com s’ha fet a la pestanya *Monotexts*— crear corpus bilingües mitjançant la càrrega de parelles de textos monolingües. Com en el cas dels monotextos, l’usuari pot combinar fitxers (sempre que siguin paral·lels) per tal d’obtenir un model de traducció més gran. El pas següent, *Translators*, permet crear traductors automàtics, amb model de llengua o sense. L’últim pas, *Translate*, permet utilitzar el motor creat, siga mitjançant la interfície web o mitjançant la càrrega de fitxers. Tal com expliquen Martín-Mor & i Huerta (2017,

<sup>12</sup>Es preveu que properament aquest pas permeti la càrrega de fitxers TMX, atés que és un format àmpliament utilitzat pels traductors. Mentrestant, es pot recórrer a programes com ara Okapi Rainbow per a la conversió de TMX a format *Parallel Corpus Files*: [http://okapiframework.org/wiki/index.php?title=Format\\_Conversion\\_Step](http://okapiframework.org/wiki/index.php?title=Format_Conversion_Step) [última visita setembre 2017].

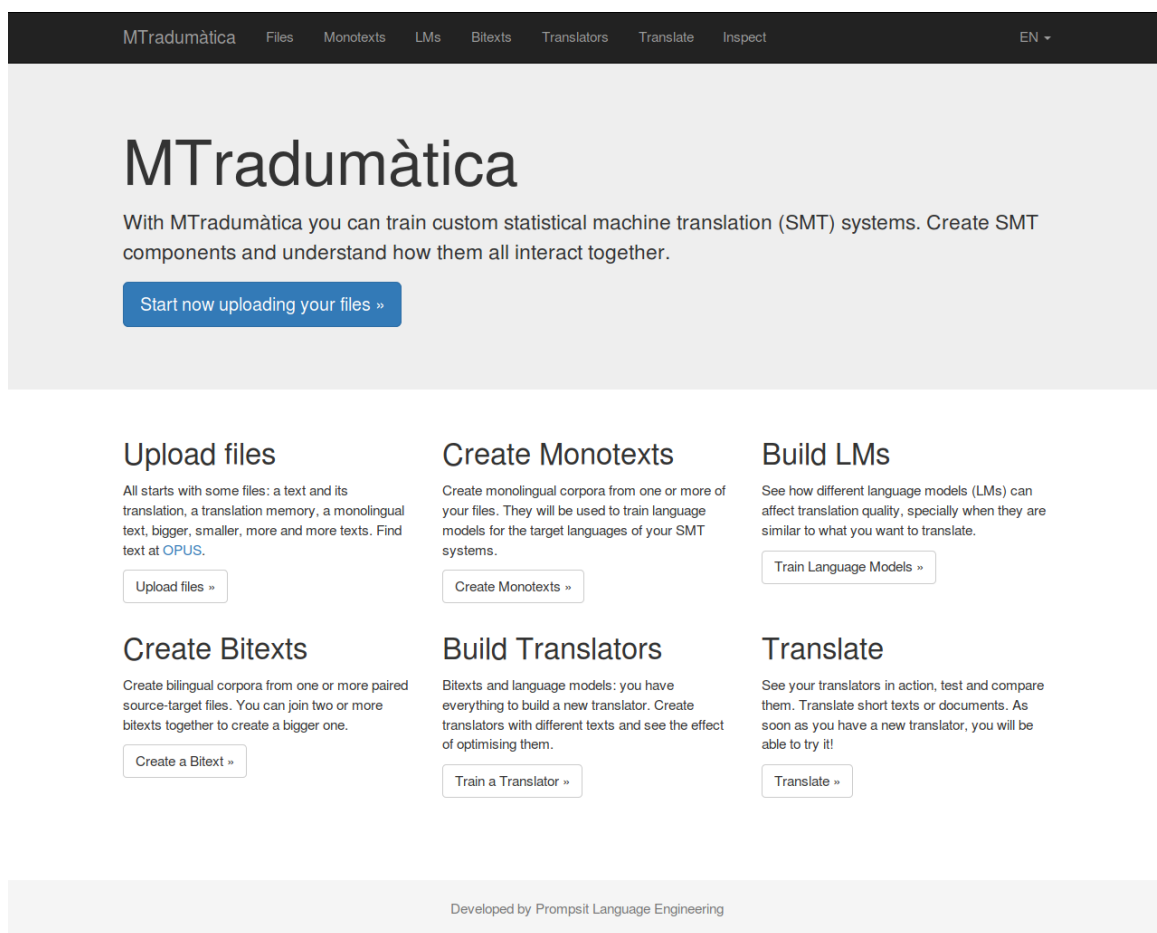


Figura 2: Interfície gràfica de MTradumàtica.

p. 112), està previst que MTradumàtica permeti a eines de traducció assistida accedir als motors mitjançant API.

Tal com s'ha esmentat anteriorment, la funció Inspect —visible en la versió actual de MTradumàtica però encara en desenvolupament— permetrà la consulta de les taules i els models de cadascun dels motors amb l'objectiu d'identificar possibles accions de millora.<sup>13</sup>

### 3 Recursos per a la creació del motor

En aquest cas pràctic d'entrenament de motors a la plataforma MTradumàtica l'objectiu ha estat crear dos motors de TAE: un del castellà al romanés i un altre del romanés al castellà. Cal precisar que, per a crear aquests motors de TAE, no és necessari emprar una plataforma com MTradumàtica, tot i que sí que facilita la tasca per la interfície gràfica i l'ús intuïtiu.

Hui en dia hi ha diversos motors de TA disponibles per a aquesta combinació de llengües,

entre els quals, el traductor de Google,<sup>14</sup> el de Yandex,<sup>15</sup> Bing de Microsoft —el qual especifica de manera explícita que utilitza l'anglès com a llengua pont—,<sup>16</sup> i el motor de TA basat en regles d'Apertium, tan sols en la direcció romanés–castellà.<sup>17</sup>

Els motors esmentats són genèrics i no estan especialitzats en cap camp del coneixement. Els motors que es presenten en aquest article han estat entrenats amb corpus especialitzats en l'àmbit de la medicina i la farmàcia, per la qual cosa han estat necessaris:

- Corpus bilingües sobre medicina i farmàcia ro↔es per a tots dos MT i tots dos ML.
- Corpus monolingüe en castellà per a millorar el ML (es).
- Corpus monolingüe en romanés per a millorar el ML (ro).

<sup>14</sup><https://translate.google.com/>.

<sup>15</sup><https://translate.yandex.com/>.

<sup>16</sup><https://www.bing.com/translator/>.

<sup>17</sup><https://www.apertium.org/>.

<sup>13</sup>Per a més informació sobre els futurs desenvolupaments, vegeu Martín-Mor (2017).



Finalment, s'han seleccionat els corpus bilingües següents:

- Corpus ro↔es de l'Agència Europea del Medicament (EMA): conté 12,9 milions de paraules en castellà i 11,9 milions en romanés, i un milió de segments alineats. El corpus naix a partir de l'alineació de fitxers en PDF d'aquest organisme i es pot descarregar de manera lliure del web del projecte Per-Fide (Almeida et al., 2014).<sup>18</sup> D'aquest corpus s'ha reservat el contingut des de la línia 961 fins a la 3461 (39 559 paraules en castellà i 36 828 en romanés) per tal de no emprar-los per a l'entrenament del motor i mantenir-los com a text de referència per a futures tasques d'optimització o per a l'avaluació automàtica de la TA (v. apartat 5). Així, s'ha entrenat el motor amb un corpus final amb 12 629 507 paraules en castellà i 11 690 520 en romanés.
- Corpus ro↔es del Centre Europeu per a la Prevenció i el Control de les Malalties (ECDC): conté 40 392 paraules en castellà i 37 105 paraules en romanés i 2 285 segments alineats. El corpus s'ha descarregat de manera lliure del portal EU Science Hub,<sup>19</sup> del servei de ciència i coneixement de la Comissió Europea.

Aquests dos corpus també s'han emprat per a entrenar els models de llengua. Tot i això, per tal de millorar aquests models, s'ha decidit entrenar-los amb continguts monolingües addicionals. Així, s'han creat dos corpus, un per a cada llengua de destí, a partir de continguts de la Viquipèdia del domini de la medicina i la farmàcia, com es veurà amb més detall a l'apartat 4. En total, els corpus monolingües tenen 378 000 paraules en castellà i 216 000 paraules en romanés.

Una de les particularitats de la combinació lingüística és la codificació dels caràcters en romanés. Les lletres diacrítics del romanés *Ș* i *Ț* (i les minúscules *ș* i *ț*) es van incloure a Unicode per primer cop al setembre del 1999, en la versió 3.0.0 (Consortium, 2000) i ISO les publica a la ISO/IEC 8859-16 un any més tard. D'altra banda, als sistemes operatius i programes no

<sup>18</sup>Vegeu <http://per-fide.di.uminho.pt/> [última visita setembre 2017].

<sup>19</sup>Vegeu <https://ec.europa.eu/jrc/en/language-technologies/ecdc-translation-memory> [última visita setembre 2017]. Els drets d'autor pertanyen a la EU/ECDC i se'n permet l'ús tant comercial com no comercial. Vegeu: [http://optima.jrc.it/Resources/ECDC-TM/2012\\_10\\_Terms-of-Use\\_ECDC-TM.pdf](http://optima.jrc.it/Resources/ECDC-TM/2012_10_Terms-of-Use_ECDC-TM.pdf) [última visita setembre 2017].

Corpus	ro	es
Viquipèdia (ro)	216 000	
Viquipèdia (es)		378 000
ECDC ro↔es	37 105	40 392
EMA ro↔es (total)	11 901 523	12 939 973
EMA ro↔es (per a l'entrenament)	11 690 520	12 629 507
EMA ro↔es (com a referència)	36 828	39 554

Taula 1: Nombre de paraules dels corpus.

s'han incorporat els caràcters correctes de manera homogènia, fet que ha provocat problemes de compatibilitat.

Això ha provocat que molts textos informatitzats en romanés no continguin diacrítics o s'hi hagen emprat durant les darreres èpoques els caràcters turcs anàlegs (*ş* i *ţ*), com descriu Kaplan (2011) amb més detall. Actualment, hi ha diversitat pel que fa a l'ús d'aquests diacrítics, el qual encara no és homogeni. Aquest és un aspecte que s'ha de tenir en compte tant durant el procés de creació del motor com durant a l'ús mateix del motor entrenat, com es veurà a l'apartat 4.

## 4 Descripció del procés de creació

En aquest apartat s'explica el procés de creació dels motors de TAE ro↔es a MTradumàtica. En primer lloc, s'hi descriuen les tasques prèvies per al processament dels corpus i, a continuació, la creació mateixa dels motors a MTradumàtica.

### Processament previ dels corpus

Com s'explica a l'apartat 2, MTradumàtica només accepta fitxers de text pla amb una sola frase per línia. Per consegüent, per a cada corpus ha sigut necessari aconseguir fitxers de text pla per a cada llengua amb una frase per línia i que conservessin l'alineació, en el cas dels corpus bilingües.

Quant al corpus de l'EMA, descarregat del web de l'OPUS, només ha calgut descarregar els fitxers en format Moses, els quals ja compleixen aquestes característiques necessàries per a la creació del motor.

Pel que fa al corpus de l'ECDC, el format per defecte és un TMX multilingüe, tot i que al paquet s'inclou un programa Java que n'extreu els parells de llengües en un TMX bilingüe. Posteriorment, tal com s'ha esmentat més amunt, s'ha convertit en fitxers en format Moses amb el programa lliure Okapi.

D'altra banda, pel que fa a la recopilació de corpus de la Viquipèdia per als ML, s'han extret articles per categories per mitjà de la funció *Exporta*.<sup>20</sup> Dins dels fitxers d'exportació, en format XML, ha calgut netejar el codi i extraure'n tan sols el text aprofitable per a entrenar el motor. Per a fer-ho s'ha emprat l'editor de textos de codi lliure *Notepad++*<sup>21</sup> mitjançant l'ús de macros. Aquest procés, inclosa la programació de les macros amb expressions regulars, s'explica amb detall a [Peña-Irles \(2017\)](#).<sup>22</sup>

Paga la pena afegir que, pels motius que s'expliquen a l'apartat 3, hi ha molts textos romanesos que no tenen diacrítics o que els tenen amb la codificació incorrecta. En el cas d'estudi de l'article no s'ha emprat cap text sense diacrítics, tot i que sí que s'ha observat heterogeneïtat dels caràcters per a l'escriptura de diacrítics romanesos. Per aquest motiu, ha calgut unificar-los. *Notepad++* ha facilitat la cerca i substitució d'aquests caràcters pels caràcters de la norma Unicode 3.0.0 de l'any 2000 ([Consortium, 2000](#)).

### Creació dels motors a MTradumàtica

Com s'explica a l'apartat 2, el primer pas és la pujada de tots els fitxers preprocessats a la plataforma MTradumàtica, tant en romanès com en castellà. Ha estat convenient modificar prèviament l'extensió dels fitxers per “.es” i “.ro” en funció de la llengua, per tal de facilitar el reconeixement de la llengua per part del sistema (v. apartat 2).

A continuació, s'ha creat un únic corpus per a cada llengua (a la pestanya *Monotexts*) a partir dels corpus previstos per als ML. Dit altrament, s'han creat dos corpus generals, un per a cada llengua, amb els fitxers d'EMEA, ECDC i els continguts de la Viquipèdia. Aquest pas és necessari per a poder completar el procés següent, és a dir, l'entrenament dels ML. A la pestanya LM, s'han entrenat dos ML de destinació, un per a cada motor de TAE, a partir dels monotextos acabats de crear. Aquest procés té una durada variable en funció de la quantitat de paraules i de la capacitat del servidor en què s'allotja MTradumàtica. En el cas del servidor de Tradumàtica, el ML en castellà s'ha entrenat en 5 minuts i 33

segons, mentre que el ML en romanès ha tardat 4 minuts i 44 segons a fer-ho.

Per als bitexts, en canvi, s'han ajuntat els corpus bilingües en un de general per tal d'entrenar els MT. En aquest cas, s'han seleccionat els corpus EMEA i ECDC, ja alineats. A diferència del procés descrit per als monotextos en el paràgraf anterior, atès que els bitextos són bidireccionals, no cal crear un corpus per a cadascun dels motors, sinó que el mateix permet entrenar tant el motor romanès–castellà com el traductor castellà–romanès.

Finalment, a la pestanya *Translators* s'han creat els traductors automàtics, un per a cada direcció, amb tot el que s'ha preparat prèviament: un ML entrenat i un bitext. Una vegada fet això, s'inicia el procés d'entrenament estadístic del motor. La durada també varia en funció de la longitud dels corpus i les especificitats del servidor en què s'allotja MTradumàtica. A tall indicatiu, en el cas del motor ro→es s'ha tardat 4 hores, 47 minuts i 43 segons, mentre que el motor es→ro ha tardat 4 hores, 46 minuts i 34 segons.

Després de l'entrenament, el pas següent per a la construcció dels motors és l'optimització. Aquest pas és optatiu i permet millorar-ne la qualitat. En el nostre cas es va ometre l'optimització dels motors pel fet que, en el moment de crear-los, aquesta funcionalitat estava en desenvolupament en MTradumàtica (vegeu l'apartat 6).

Un cop completat l'entrenament, a la pestanya *Translate* els traductors automàtics creats ja es poden utilitzar, tant introduint-hi un text a la interfície web, com mitjançant la càrrega de fitxers. Per a la combinació romanès–castellà cal tenir en compte que és necessari que els textos tinguin els caràcters de l'Unicode 3.0.0 abans d'introduir un text per a traduir-lo. Altrament, el traductor no és capaç de reconèixer els caràcters no estandarditzats, ja que no apareixen als corpus amb què s'ha entrenat el motor.

## 5 Resultats

L'objectiu d'aquest apartat és analitzar el rendiment dels motors mitjançant mètriques automàtiques per tal de demostrar la viabilitat d'MTradumàtica per a l'entrenament dels motors i comparar aquestes mètriques amb altres motors de TAE. S'han avaluat els resultats dels motors de traducció mitjançant tres mètriques d'avaluació automàtica de la TA: BLEU ([Papineni et al., 2002](#); [KantanMT](#)), METEOR-ex ([Banerjee & Lavie, 2005](#)) i TER ([Snover et al., 2006](#)). Els dos primers mètodes es mesuren mitjançant valors de l'1 al 0, i n'és l'1 el valor òptim segons

<sup>20</sup>La funció *Exporta* permet la descàrrega d'articles per categories. Disponible a l'adreça <https://ro.wikipedia.org/wiki/Special:Export%C4%83> [última visita setembre 2017].

<sup>21</sup>Aquest editor de textos es pot descarregar des del web <https://notepad-plus-plus.org/> [última visita setembre 2017].

<sup>22</sup>El fitxer de macros ha estat publicat amb llicència lliure a <http://www.github.com/tradumatica>.

aquest mètode. D'altra banda, el mètode TER és un indicador que mesura l'esforç de postedició, de manera que, com més baix és el valor, menor és l'esforç de postedició (Peña-Irles, 2017). Aquests mètodes són avaluadors automàtics que indiquen el rendiment del motor, tot i que no expressen necessàriament la qualitat del resultat de la TA.

Per a dur a terme les avaluacions automàtiques és necessari disposar d'un text original, una o diverses traduccions automàtiques i una o diverses traduccions amb qualitat humana. S'ha fet servir el conjunt d'eines d'avaluació automàtica de la TA anomenat *Asiya Online*,<sup>23</sup> desenvolupat per la Universitat Politècnica de Catalunya, un “open toolkit aimed at covering the evaluation needs of system and metric developers along the development cycle” (Giménez & Màrquez, 2010). És accessible de manera lliure pel web i permet valorar els resultats de la TA mitjançant més de quinze mètodes d'avaluació (Peña-Irles, 2017). Cal precisar que *Asiya Online* mostra l'indicador TER en valors negatius (–TER), per la qual cosa n'ha estat necessària la conversió a valors positius.

Pel que fa a l'avaluació dels motors entrenats, se n'han dut a terme dues per a cada combinació de llengües, la primera a partir d'un text de referència extret del corpus de l'EMEA,<sup>24</sup> i la segona a partir d'un prospecte mèdic posteditat per a cada llengua.<sup>25</sup> D'altra banda, s'han dut a terme les mateixes avaluacions amb tres altres motors de TA genèrics disponibles en aquestes combinacions d'idiomes: Google, Yandex i Apertium —només per a la combinació romanés–castellà— (v. l'apartat 3), per tal de comparar-ne els resultats. Recordem també (vegeu l'apartat 4, Descripció del procés de creació) que els motors que s'analitzen a continuació no han estat optimitzats. La metodologia i els resultats d'aquesta anàlisi s'analitzen amb més detall a Peña-Irles (2017).

<sup>23</sup>L'URL de l'eina és [http://asiya.lsi.upc.edu/demo/asiya\\_online.php](http://asiya.lsi.upc.edu/demo/asiya_online.php) [última visita setembre 2017].

<sup>24</sup>El text de referència s'ha pres d'una part d'un corpus que s'ha extret prèviament a l'entrenament del motor i que s'empra amb la finalitat d'avaluar el rendiment del motor i per a l'ajustament dels paràmetres de la TAE o *tuning*. Els textos emprats per a l'avaluació tenen 2 203 paraules en castellà i 2 083 en romanés.

<sup>25</sup>Per al castellà s'ha descarregat un prospecte del web del Ministeri de Sanitat espanyol (de 185 paraules): [https://www.aemps.gob.es/cima/dohtml/p/69429/Prospecto\\_69429.html](https://www.aemps.gob.es/cima/dohtml/p/69429/Prospecto_69429.html) [última visita setembre 2017]. Per al romanés, s'ha extret un prospecte del web *Ce se întâmplă doctore* (de 269 paraules): <http://www.csid.ro/medicamente/omeprazol-terapia-20-mg-capsule-gastrorezistente-11474561/> [última visita setembre 2017].

## Motor romanés–castellà

Els resultats del motor entrenat a MTradumàtica del romanés al castellà amb les mètriques esmentades es mostren a la taula següent. També s'hi comparen els resultats amb els dels traductors d'Apertium, Google i Yandex:

	BLEU	METEOR-ex	TER
<b>Text de referència EMEA (ro→es)</b>			
MTradumàtica	0,60	0,73	0,35
Apertium	0,19	0,35	0,68
Google	0,43	0,58	0,47
Yandex	0,35	0,54	0,54
<b>Prospecte posteditat (ro→es)</b>			
MTradumàtica	0,54	0,69	0,32
Apertium	0,33	0,51	0,45
Google	0,40	0,59	0,35
Yandex	0,52	0,66	0,29

Taula 2: Avaluació de la TA ro→es (Mtradumàtica, Apertium, Google i Yandex).

La taula anterior mostra que els resultats de les mètriques d'avaluació amb MTradumàtica són similars, i fins i tot, en la majoria dels casos, superiors, als de productes existents, la qual cosa indica que el motor descrit en aquest article podria ser viable en aplicacions de disseminació (Forcada, 2009). Els resultats del motor romanés–castellà presenten unes mètriques molt positives i elevades, que podrien suposar la viabilitat del motor en aplicacions de disseminació. D'altra banda, en comparar-lo amb la resta de motors de TA analitzats, s'obtenen uns resultats superiors. Hi destaca el resultat de Google, a la taula 3, amb diferències d'entre 0,15 i 0,2 punts al paràmetre BLEU, i el baix rendiment d'Apertium. A més, el resultat de Yandex en el text posteditat és semblant i, fins i tot, superior en el cas de TER.

## Motor castellà–romanés

Els resultats del motor entrenat a MTradumàtica del castellà al romanés amb les mètriques esmentades es mostren a la taula següent. També s'hi comparen els resultats amb els dels traductors de Google i Yandex:

En analitzar els resultats per a la combinació castellà–romanés s'observa que les mètriques són molt positives i que, a més a més, s'obtenen els millors resultats en comparació amb els altres dos motors analitzats. Els resultats tant de Google com de Yandex obtenen unes mètriques inferiors.

	BLEU	METEOR-ex	TER
<b>Text de referència (es→ro)</b>			
MTradumàtica	0,73	0,51	0,24
Google	0,33	0,30	0,54
Yandex	0,29	0,28	0,58
<b>Prospecte posteditat (es→ro)</b>			
MTradumàtica	0,54	0,47	0,34
Google	0,35	0,30	0,44
Yandex	0,30	0,29	0,49

Taula 3: Avaluació de la TA es→ro (Mtradumàtica, Apertium, Google i Yandex).

## 6 Conclusions

Aquest article ha presentat un estudi de cas d'aplicació d'un producte de recerca a un projecte real. La plataforma MTradumàtica, desenvolupada en el marc d'un projecte públic per a l'acostament de la traducció automàtica als traductors, i amb un èmfasi en l'aspecte formatiu, ha estat utilitzada per part de traductors per crear un traductor automàtic especialitzat en farmàcia i medicina per a la combinació lingüística romanès–castellà. D'una banda, això ens permet constatar que l'objectiu per al qual naixia el producte en certa manera es compleix: s'ha creat un motor de TAE especialitzat a partir de recursos lliures de la xarxa utilitzant la interfície gràfica de la plataforma. El motor de TAE, a més, no sols és funcional, sinó que dona bons resultats pel que fa al rendiment amb indicadors aproximats com BLEU, com es veu a l'apartat 5. Cal tenir en compte, com hem dit a l'apartat 4, que les mètriques automàtiques presentades en aquest article s'han generat a partir de motors no optimitzats. És bastant raonable pensar que els motors optimitzats obtindrien millors resultats, per la qual cosa podem considerar que els valors obtinguts són un bon punt de partida que només podria millorar. Malgrat tot, creiem que l'interès de l'experiència es troba no tant en el rendiment del resultat, sinó principalment en la constatació que és possible dur a terme processos de creació de motors de qualitat per part de traductors. En aquest sentit, com a línia de recerca en un futur, seria interessant analitzar la qualitat real dels resultats d'aquests motors optimitzats amb indicadors de l'esforç de postedició, com ara HTER, mitjançant experiments reals de posteditors. D'altra banda, l'experiència descrita suggereix, tal com plantejava el projecte de recerca de Tradumàtica, que el perfil dels traductors és un perfil vàlid per dur a terme processos relacionats amb la personalització de motors de TA.

Tot això ens fa entreveure l'impacte que pot tenir per als programes de formació en traducció el desenvolupament de sistemes de personalització de TA amb interfície gràfica.

L'article ha descrit detalladament cadascun dels passos que s'han seguit per al desenvolupament d'un motor amb l'objectiu que l'experiència siga replicable per part d'altres traductors amb necessitats similars, per a la mateixa o per a altres combinacions lingüístiques i camps d'especialitat. És per aquest motiu que s'ha utilitzat no sols programari lliure sinó també recursos disponibles amb llicències lliures. També els recursos generats, com ara el paquet de macros per a la neteja dels corpus descarregats de la Viquipèdia han estat posats a disposició de la comunitat amb llicència lliure.

L'experiència descrita apunta a la necessitat que les plataformes per a la personalització de motors permeten el preprocessament dels corpus mitjançant regles senzilles. En casos com l'esmentat a l'apartat 3, seria útil configurar una sèrie de regles, com ara per mitjà de la utilitat d'Unix *Stream EDitor (sed)*,<sup>26</sup> amb llicència GPLv3, útil per a aplicar transformacions a un text, amb l'objectiu que qualsevol codificació incorrecta en el text original no genere una traducció errònia o desconeguda, sinó que es convertisca a la codificació correcta abans de ser traduïda.

## Referències

- Almeida, José João, Sílvia Araújo, Nuno Carvalho, Idalete Dias, Ana Oliveira, André Santos & Alberto Simões. 2014. The Per-Fide corpus: A new resource for corpus-based terminology, contrastive linguistics and translation studies. En *Working with Portuguese Corpora*, 177–200. Bloomsbury Publishing.
- Banerjee, Satanjeev & Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. En *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
- Consortium, Unicode. 2000. *The unicode standard, version 3.0*. Addison-Wesley.
- Forcada, Mikel L. 2009. Apertium: traducció automàtica de codi obert per a les llengües romàniques. *Linguamàtica* 1(1). 13–23.
- Giménez, Jesús & Lluís Màrquez. 2010. Asiya: An open toolkit for automatic machine trans-

<sup>26</sup>Vegeu <http://sed.sourceforge.net/sedfaq2.html#s2.1> [última visita setembre 2017].



- lation (meta-)evaluation. *The Prague Bulletin of Mathematical Linguistics* 94(1). 77–86.
- KantanMT. 2017. What is BLEU score? Recuperat de <https://www.kantanmt.com/whatisbleuscore.php>.
- Kaplan, Michael S. 2011. The history of messing up Romanian on computers. MSDN blogs. 24 d'agost. Recuperat de <http://archives.miloush.net/michkap/archive/2011/08/24/10199324.html>.
- Koehn, Philipp. 2016. *Moses user manual and code guide*.
- LT-Innovate. 2013. Status and potential of the European language technology markets. [http://ec.europa.eu/information\\_society/newsroom/cf/dae/document.cfm?doc\\_id=4267](http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=4267).
- Machado, Maria José & Hilário Leal Fontes. 2014. *Moses for mere mortals. tutorial. a machine translation chain for the real world*. <https://github.com/jladcr/Moses-for-Mere-Mortals/blob/master/Tutorial.pdf>.
- Martín-Mor, Adrià & Ramon Piqué i Huerta. 2017. MTradumàtica i la formació de traductors en traducció automàtica estadística. *Tradumàtica* 15. 97–115.
- Martín-Mor, Adrià. 2017. MTradumàtica: Statistical machine translation customisation for translators. *Skase Journal of Translation and Interpretation* 11(1). 25–40.
- Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. En *40th annual meeting on Association for Computational Linguistics (ACL'2002)*, 311–318.
- Peña-Irles, Víctor. 2017. Entrenament de motors de traducció automàtica estadística especialitzats en farmàcia i medicina entre el castellà i el romanés. Treball de recerca de màster. Universitat Autònoma de Barcelona.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciula & John Makhoul. 2006. A study of translation edit rate with targeted human annotation. En *Association for Machine Translation in the Americas Conference*, s.pp.
- Tiedemann, Jörg. 2009. News from OPUS: A collection of multilingual parallel corpora with tools and interfaces. En *Recent Advances in Natural Language Processing*, 237–248. John Benjamins.
- Torres-Hostench, Olga, Marisa Presas & Pilar Cid-Leal. 2016. L'ús de traducció automàtica i postedició a les empreses de serveis lingüístics de l'estat espanyol. informe de recerca ProjeccTA 2015. Universitat Autònoma de Barcelona.