MTradumàtica i la formació de traductors en Traducció Automàtica Estadística

Adrià Martín-Mor, Ramon Piqué
Grup Tradumàtica

Adrià Martín-Mor
adria.martin@uab.cat

Ramon Piqué
ramon.pique@uab.cat

Resum

El present article es proposa acostar-se a la formació de traductors en l'àmbit de la traducció automàtica estadística (TAE), específicament des de l'òptica de la personalització de motors. Basant-se en l'informe de ProjecTA del grup Tradumàtica, s'identifica la formació com un dels factors imprescindibles a l'hora d'acostar la traducció automàtica (TA) als professionals de la traducció, siguin traductors professionals o empreses de traducció. A partir d'aquest diagnòstic, elaboren una proposta docent modular (que es pot segmentar en diverses activitats) adaptable a diversos contextos formatius en el camp de la traducció. Alhora es descriu la versió experimental de la plataforma web MTradumàtica com una eina per a la formació i la recerca en l'àmbit.

Paraules clau: traumàtica; traducció automàtica; formació; postedició; personalització de motors de traducció automàtica

Resumen

El presente artículo tiene el propósito de acercarse a la formación de traductores en el ámbito de la traducción automática estadística (TAE), específicamente desde la óptica de la personalización de motores. Basándose en el informe de proyectos del grupo Tradumàtica, se identifica la formación como uno de los factores imprescindibles a la hora de acercar la traducción automática (TA) los profesionales de la traducción, sean traductores profesionales o empresas de traducción. A partir de este diagnóstico, elaboran una propuesta docente modular (que se puede segmentar en diversas actividades) adaptable a diversos contextos formativos en el campo de la traducción. Asimismo se describe la versión experimental de la plataforma web MTradumàtica como una herramienta para la formación y la investigación en el ámbito.

Palabras clave: tradumática; traducción automática; formación; posedición; personalización de motores de traducción automática.

Abstract

This article proposes bringing translator training closer to the sphere of statistical machine translation (SMT),

Rebuda: 21 de juliol de 2017 / Acceptació: 16 de desembre de 2017 / Publicació 29 de desembre de 2017

particularly from the perspective of personalising them. Based on the Tradumàtica research group ProjecTA report, training sessions are identified as one of the essential factors when introducing professional to automatic translation, whether individual professional translators or translation companies. Based on this diagnostic, a modular teaching proposal is proposed, which can be broken down into various activities, and adapted to different training contexts in the field of translation. Along these lines, it describes the experimental version of the MTradumàtica web-based platform as a tool for training and research in this area.

Keywords: tradumàtica; automatic translation; training; post-editing; customization of machine translation engines.

1. Introducció[1]

La irrupció de la informàtica en l'àmbit de la traducció ha provocat canvis en l'activitat professional d'una magnitud que fan difícil de reconèixer aquesta activitat en relació amb la que va ser exercida en èpoques recents pels mateixos professionals. En relativament pocs anys el perfil del traductor professional s'ha vist profundament sacsejat pels avenços tecnològics, tant pel que fa a l'objecte de la traducció com pel que fa a les eines amb què traduïm. Actualment assistim a la internetització del procés de traducció —un estadi caracteritzat pel fet que cada cop més les tasques es duen a terme en línia (Martín-Mor, Piqué i Sánchez-Gijón, 2016: 23)— i a la consolidació de la postedició (PE) com a tasca nuclear en l'acció traductora. En aquest sentit es pot apuntar que estem vivint un canvi de paradigma a molts nivells de la professió que comporta també canvis de paradigma pel que fa a la competència instrumental i conseqüentment a la formació del traductor.

1 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.

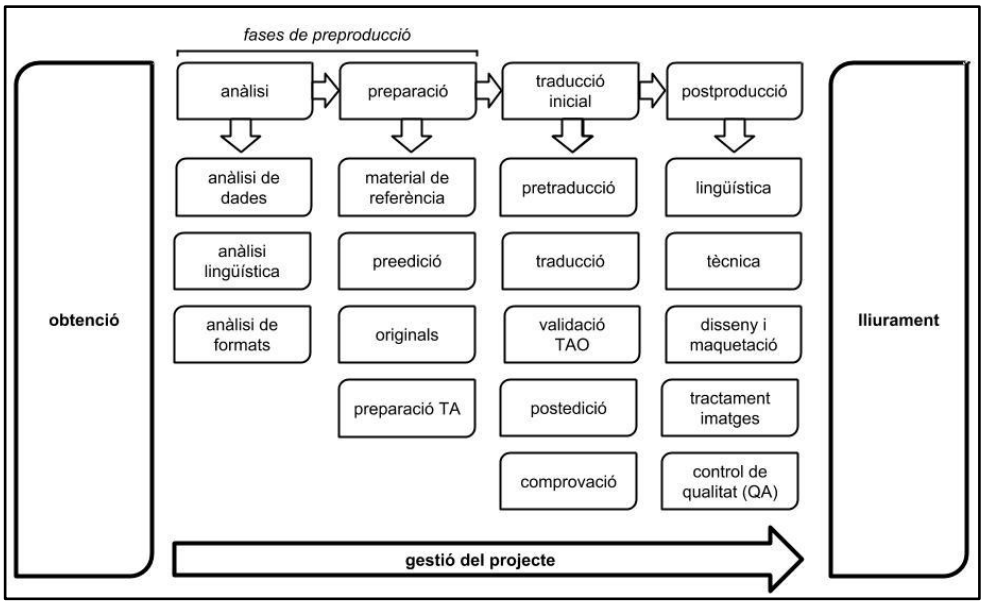


Figura 1: La digitalització del procés de traducció (Martín-Mor, Piqué i Sánchez-Gijón, 2016).

La figura 1 descriu el procés de digitalització de la traducció i mostra, organitzades en fases, les diferents tasques que ha de dur a terme un traductor en un projecte tipus de traducció professional. En aquest article volem posar l'accent sobre les tasques relacionades amb la traducció automàtica estadística (TAE) i la PE com a esquema de treball emergent en l'àmbit professional.

L'estudi de ProjecTA[2] sobre el grau d'implementació de la traducció automàtica (TA) i la PE en les empreses (Torres-Hostench, Presas i Cid-Leal, 2016) posava en relleu el fet que gairebé la meitat de les empreses enquestades feia servir la TA en el flux de treball, malgrat que una part rellevant només el feia servir en un 10% dels projectes de traducció. Una altra dada destacable era que només un 16% feia servir un sistema propi de TA. Ambdues dades són remarcades en l'informe del projecte com a mereixedores d'un estudi més aprofundit.

Una de les conclusions de l'informe té relació amb el desconeixement de les noves tasques derivades de la implementació de la TA en el flux de treball per part del professional de la traducció. Aquest desconeixement «pot generar inseguretat en els professionals i fins i tot provocar reaccions de rebuig», mentre que, per contra, si el traductor rep una formació adequada «podrà gestionar les fases del procés amb TA com ara la preparació de la documentació, l'entrenament del sistema, la preedició de textos, la postedició i la retroalimentació del sistema» (Torres-Hostench, Presas i Cid-Leal, 2016: 1).

Resultats de ProjecTA són també les sessions del grup focal que es va reunir amb diverses empreses de traducció per discutir a l'entorn de diferents qüestions relacionades amb l'ús de la TA en les empreses. De la riquesa de les respostes i dels debats associats, en remarquem l'asseveració compartida pels assistents que la «TA és

[2] http://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.

una evolució de la traducció» (p. 36) i que la manca de coneixements és per moltes empreses un dels reptes a l'hora d'implementar sistemes de TA propis (p. 37).

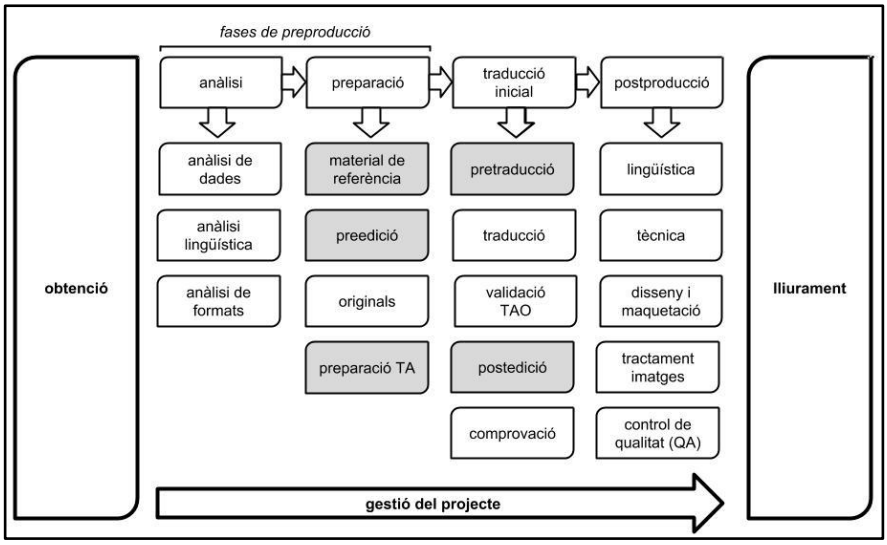


Figura 2: Tasques relacionades amb l'ús de la TA i la PE (Martín-Mor, Piqué i Sánchez-Gijón, 2016).

En l'esquema del procés digitalitzat de la traducció, les tasques objecte del treball relacionades amb la TA i la PE són les que s'apunten a la figura 2: la preparació del material de referència, la preedició, la pretraducció i la postedició, ubicades dins les categories de preparació i de traducció.

L'informe ProjecTA (p. 28) recull de manera explícita que

> a llarg termini la implantació d'un sistema adaptat amb corpus propis milloraria la productivitat. També sembla que a les empreses els falta capacitat tecnològica i econòmica per a implantar un sistema de TA propi. Això ens porta a pensar que en el futur es podria produir una gran diferència entre les empreses en funció de les seves capacitats tecnològiques.

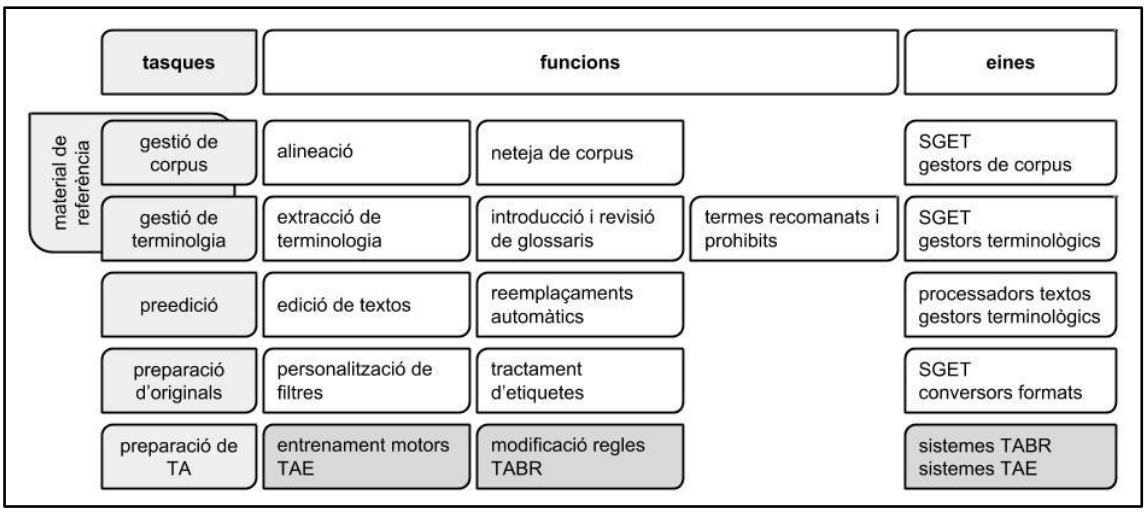


Figura 3: Tasques de la fase de preparació (Martín-Mor, Piqué i Sánchez-Gijón, 2016).

En l'article volem posar l'accent sobre la preparació TA, especialment en l'entrenament de motors de TAE, tal com s'indica en la figura 3, deixant de banda les

tasques relacionades amb la preedició (que caldrà abordar en un altre text i en una altra ocasió, ja que queda fora de la recerca de ProjecTA). Per aquesta mateixa raó ens centrarem en l'entrenament de motors de TAE, tecnologia sobre la qual es va desenvolupar una part del projecte apuntat.

2. Personalització de motors de TAE

Avui, hi ha algunes plataformes que permeten la personalització de motors de TAE. KantanMT, [3] LetsMT, [4] Microsoft Translator Hub [5] o Slate Desktop (anteriorment, DoMosesYourself) [6] són bons exemples de l'àmbit del programari privatiu. Moses, [7] programari lliure amb llicència GNU Lesser General Public License, és un dels sistemes més utilitzats per a la creació de motors de TAE: «[Moses is] widely used within the industry to build customized MT engines» (LT-Innovate, 2013: 71). 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).» Amb tot, tal com fa notar LT-Innovate (2013: 72), Moses és difícil d'administrar, entre altres coses perquè no disposa d'una interfície gràfica d'usuari (GUI). Aquest sol fet, que implica que l'usuari ha de tenir coneixements de sistemes UNIX i del terminal, ja comporta una barrera per a molts dels usuaris potencials. És per aquest motiu que els últims anys han sorgit diversos intents per a acostar els sistemes de TA a un públic menys expert. Machado i Leal Fontes (2014), per exemple, van desenvolupar un paquet de programes lliures («by a translator for translators», p. 2) per a la creació de motors de TAE, incloent-hi eines per convertir formats o materials de suport. Paral·lelament, han anat naixent sistemes basats en Moses amb interfície gràfica i llicències lliures, com ara ModernMT,[8] Machine Translation Training Tool (MTTT),[9] o MTradumàtica.[10]

MTradumàtica, 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 naixia amb els propòsits següents:

1/ Desenvolupar una interfície gràfica prenent en consideració una dimensió educativa envers l'usuari final.

[3] KantanMT <https://www.kantanmt.com/>
[4] LetsMT <https://www.letsmt.eu/>
[5] Microsoft Translator Hub <https://hub.microsofttranslator.com/>
[6] Slate Desktop <https://slate.rocks/>
[7] Moses <https://www.statmt.org/moses/>
[8] ModernMT <https://www.modernmt.eu/>
[9] Machine Translation Training Tool <https://github.com/roxana-lafuente/MTTT/>
[10] MTradumàtica <https://m.tradumatica.net/>

2/ Permetre l'ús via web, per tal de poder prescindir d'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 un major grau de 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.

El projecte de desenvolupament de la plataforma MTradumàtica en el marc de ProjecTA contemplava els processos esmentats en Moses per a l'entrenament de motors de TAE i també la incorporació d'un mòdul de PE que hauria de permetre retroalimentar el model de traducció.

L'objectiu de l'eina és proporcionar a investigadors no experts en tecnologia una aplicació web que permeti crear un motor de traducció amb Moses. L'aplicació vol servir, també, com a prova de concepte per a les empreses de traducció i els posteditors que vulguin posar a prova un flux de treballa amb motors propis de TA, sense oblidar el vessant didàctic en el marc de la docència de processos de TA adreçada a estudiants de traducció.

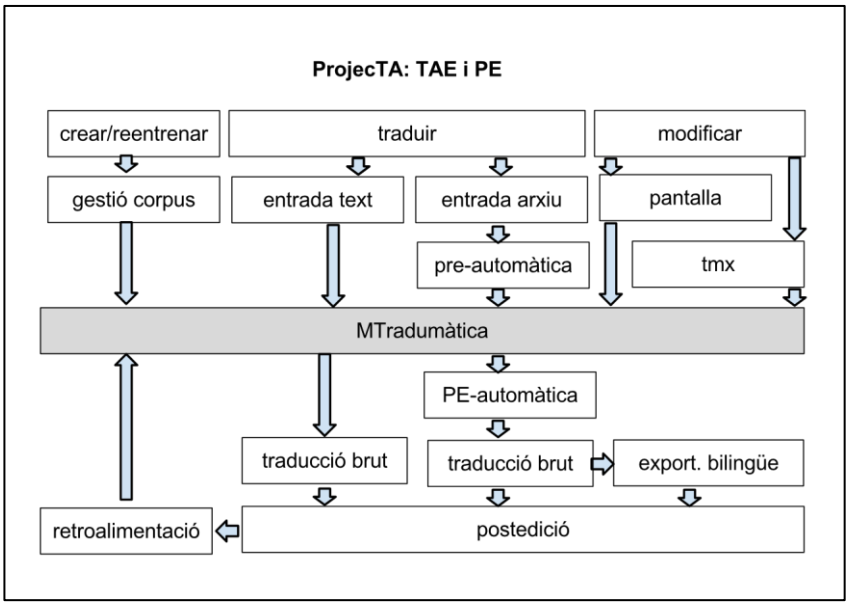


Figura 4: Plataforma TAE+PE.

En la primera fase del projecte les tasques desenvolupades i implementades en MTradumàtica han estat les que es poden observar en la figura 5.

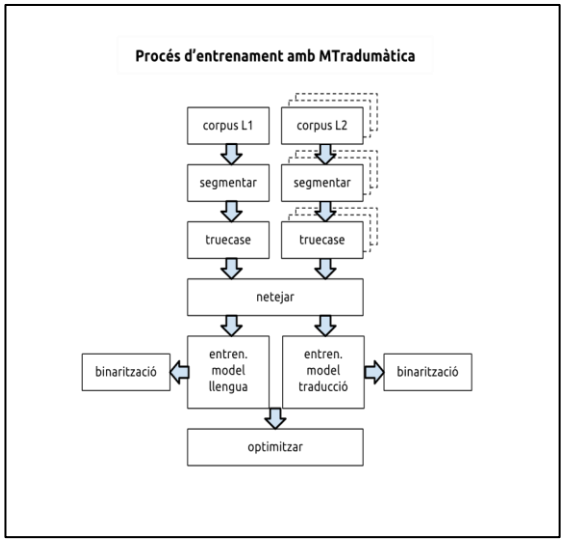


Figura 5: Esquema de procés d'entrenament en MTradumàtica.

El procés de treball amb MTradumàtica parteix 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: 36), duu a terme els processos de segmentació, *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. 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. Tot seguit 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. Atès que la consulta de les taules pot ser lenta, els models es binaritzen per tal que es carreguin 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: 12). Durant el procés d'optimització, es tradueixen automàticament milers de frases d'un subconjunt dels models (anomenat *development* o *tuning set*), es comparen amb les traduccions humanes de referència i s'ajusten automàticament els valors de cada paràmetre per tal de millorar la qualitat del motor, mesurada mitjançant mètriques automàtiques com ara BLEU (Papineni, Roukos, Ward, i Zhu, 2002).

La interfície del programa reflecteix l'objectiu principal de ProjecTA —l'acostament de la TA als traductors—, per la qual cosa hi apareixen referències als processos esmentats anteriorment. La intencionalitat didàctica de l'eina, doncs, és evident, per tal que fins i tot els usuaris que no tenen cap coneixement sobre creació de motors de TAE puguin fer-la servir i aprendre les nocions bàsiques de l'àmbit.

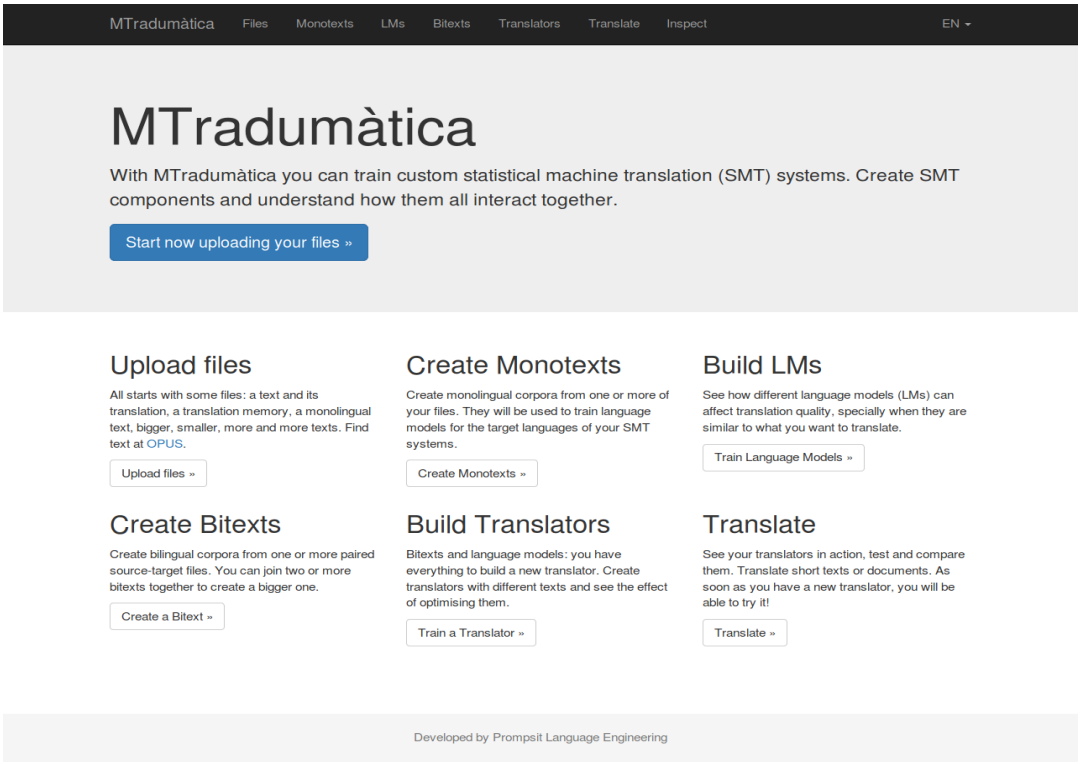


Figura 6: Interfície gràfica de MTradumàtica.

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

2/ Càrrega de fitxers

3/ Creació de monotextos

4/ Generació de models de llengua

5/ Creació de bitextos

6/ Generació de traductors automàtics

7/ Traducció

La pàgina inicial presenta els sis passos, juntament amb una breu explicació i indicacions addicionals. Al llarg de tot el procés, la barra superior mostra a l'usuari en quin pas es troba.

La creació d'un motor comença amb la càrrega dels fitxers a partir dels quals es generaran els models de llengua i de traducció. Per tal de facilitar l'obtenció de corpus, la pàgina inicial inclou un enllaç al projecte Opus, el repositori de corpus lliures (Tiedemann, 2009). Els textos carregats (un per a cada llengua, com a mínim) es mostren a la pestanya Files, juntament amb informació quantitativa (nombre de línies, paraules i caràcters) i la llengua del fitxer, detectada automàticament pel programa (l'usuari té la possibilitat de modificar-la en els casos en què la detecció automàtica falli). A sota dels textos carregats, hi ha un camp per a la càrrega de

fitxers. Tal com s'informa al peu de la pestanya Files, actualment només es poden carregar fitxers de text amb una sola frase per línia.11

La pestanya Monotexts permet combinar fitxers monolingües carregats anteriorment amb l'objectiu que, al pas següent, es puguin generar models de llengua des de la pestanya LMs. Com més fitxers es combinin, més gran serà el model de llengua. Un procés similar al dels monotextos tindrà lloc a continuació, aquest cop amb bitextos (parelles de textos original-traducció), amb l'objectiu de generar models de traducció (MT). La pestanya Bitexts permet també combinar diversos fitxers (sempre que siguen paral·lels) per MT més grans. El pas següent, Translators, permet crear traductors automàtics, amb model de llengua o sense. L'últim pas, Translate, permet utilitzar el motor creat, sigui mitjançant la interfície web o mitjançant la càrrega de fitxers.

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.

3. L'entrenament de motors en la formació de traductors

Sánchez-Gijón (2016: 157) aborda quin ha de ser el perfil del posteditor preparat per a gestionar el procés de traducció, i posa en relleu la capacitat de presa de decisions sobre la idoneïtat del recursos. L'autora emfatitza que "[e]n cuanto a la preparación de motores o modificación de sistemas de TA, se trata de tareas vinculadas a la creación u optimización de motores de TA estadística, o bien la modificación y mejora de sistemas de TA basados en reglas".

Les activitats de formació relacionades amb l'adquisició de la competència instrumental en l'entrenament de motors de TA haurien de contemplar tasques relacionades amb la gestió del material de referència —principalment el tractament de corpus monolingües i bilingües—, prèvies a l'entrenament amb MTradumàtica. La figura 7 recull el detall de les diferents tasques amb una indicació de les eines apropiades. Essent els corpus la primera matèria per a l'entrenament de sistemes, cal considerar bàsiques tasques i eines relacionades amb el tractament de textos, des de la cerca i la descàrrega dels corpus fins a la neteja i la conversió de formats. Entre les eines que poden resultar útils per a aquest propòsit, hi ha metacercadors, gestors de corpus, navegadors fora de línia o eines especialitzades com Bicrawler.12 També l'alineació de fitxers és una de les tasques recurrents a l'hora de crear corpus bilingües. MTradumàtica permet entrenar motors i dur a terme processos de traducció; actualment, la postedició de les traduccions no es pot dur a terme des de la plataforma, i cal recórrer o bé a eines específiques (com ara Post-Editing Tool; v. Aziz,

[11] Mentre no s'implementi en MTradumàtica una funció per a la conversió del conegut estàndard TMX a format Moses, es pot recórrer a programes com Okapi Rainbow: https://okapiframework.org/wiki/index.php?title=Format_Conversion_Step [última visita: 14 de desembre del 2017]. Vegeu també 4.2, 4.2. Obtenció de corpus paral·lels en TMX i conversió a format Moses.

[12] Bicrawlert <https://bicrawler.com/>

Castilho i Specia, 2012) o bé a eines SGET (vegeu 4.5, 4.5. Activitat complementària — postedició de TA).

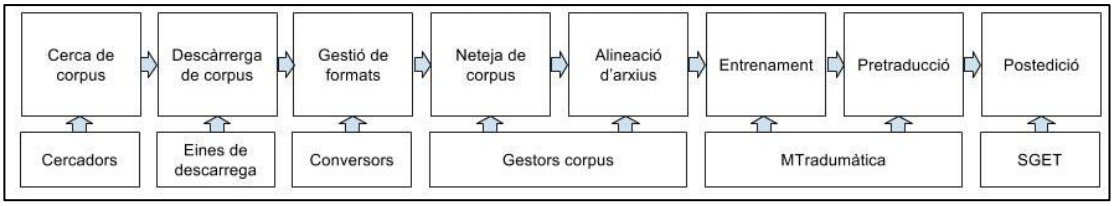


Figura 7: Tasques i eines del procés d'entrenament de motors TAE

4. Proposta docent per a la creació de motors de TA personalitzats

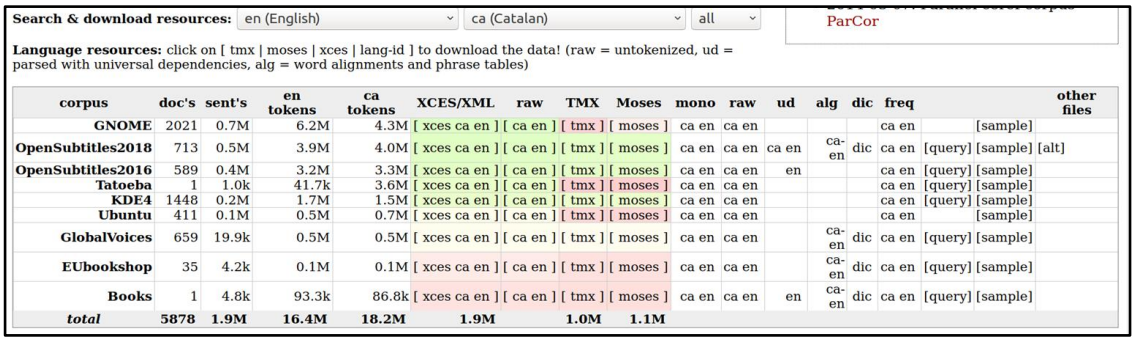
El que segueix és una proposta d'activitat docent relacionada amb la creació de motors de TA personalitzats. L'activitat ha estat duta a terme parcialment o completa en diversos programes de formació per a traductors (tant de grau com de màster) relacionats amb el grup de recerca Tradumàtica i en diverses combinacions lingüístiques (anglès-català; anglès-sard). L'activitat completa ha estat posada a prova i duta a terme en set hores de classe, malgrat que, atesa la modularitat de les parts que la componen, es pot allargar o escurçar en funció de les característiques del curs.

La proposta docent té per objectiu la creació d'un motor de TA des de zero amb MTradumàtica, i preveu l'obtenció de corpus paral·lels des de diversos recursos, la conversió de TMX a format Moses, la descàrrega i neteja de corpus monolingües, la generació de models i la postedició de TA.

Cadascuna de les seccions següents conté una tasca de l'activitat, amb referències als conceptes teòrics associats i als processos tècnics que cal dur a terme.

4.1. Obtenció de corpus paral·lels d'Opus i càrrega a MTradumàtica

L'activitat comença amb la descàrrega d'un corpus paral·lel del repositori d'Opus (vegeu 2, 2. Personalització de motors de TAE). En aquesta primera tasca, el docent pot introduir conceptes relacionats amb la lingüística de corpus, com ara *corpus paral·lel* o *bitext*. En funció de la combinació lingüística seleccionada, es poden analitzar els diversos recursos allotjats a Opus, com ara el Diari Oficial de la Generalitat de Catalunya (DOGC, català-castellà), el diari de sessions del parlament europeu (EUROPARL) o de les Nacions Unides (MultiUN), i també recursos creats per comunitats de traducció (OpenSubtitles, Viquipèdia o traduccions de productes de programari lliure com Ubuntu o GNOME). Des del punt de vista tècnic, es pot fer referència als diversos sistemes que Opus fa servir per a la codificació de les llengües: el codi de llengua (com ara *ca* per al català) o la combinació llengua-regió (*zh_TW* per al xinès de Taiwan). Un cop seleccionades les llengües del projecte per mitjà dels desplegables de la pàgina inicial d'Opus, la informació dels corpus es mostrarà en una taula com la següent:



Search & download resources: en (English) | ca (Catalan) | all

ParCor

Language resources: click on [tmx | moses | xces | lang-id] to download the data! (raw = untokenized, ud = parsed with universal dependencies, alg = word alignments and phrase tables)

corpus	doc's	sent's	en tokens	ca tokens	XCES/XML	raw	TMX	Moses	mono	raw	ud	alg	dic	freq		other files
GNOME	2021	0.7M	6.2M	4.3M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en				ca en	[sample]	
OpenSubtitles2018	713	0.5M	3.9M	4.0M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en	ca en	ca-en	dic	ca en	[query] [sample]	[alt]
OpenSubtitles2016	589	0.4M	3.2M	3.3M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en	en			ca en	[query] [sample]	
Tatoeba	1	1.0k	41.7k	3.6M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en				ca en	[query] [sample]	
KDE4	1448	0.2M	1.7M	1.5M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en				ca en	[query] [sample]	
Ubuntu	411	0.1M	0.5M	0.7M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en				ca en	[sample]	
GlobalVoices	659	19.9k	0.5M	0.5M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en		ca-en	dic	ca en	[query] [sample]	
EUbookshop	35	4.2k	0.1M	0.1M	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en		ca-en	dic	ca en	[query] [sample]	
Books	1	4.8k	93.3k	86.8k	[xces ca en]	[ca en]	[tmx]	[moses]	ca en	ca en	en	ca-en	dic	ca en	[query] [sample]	
total	5878	1.9M	16.4M	18.2M	1.9M		1.0M	1.1M								

Figura 8: Descàrrega de corpus amb Opus.

La taula mostra el nombre de documents, frases, paraules (en llengua de partida i d'arribada) i enllaços de descàrrega en diversos formats, entre els quals TMX i Moses. Aquest últim enllaç permetrà descarregar un paquet comprimit en format zip amb dos fitxers, el de la llengua de partida i el de la llengua d'arribada. Els fitxers estaran codificats amb tres elements separats per punts: una referència al subcorpus en concret (com ara *opensubtitles*), la combinació lingüística separada per un guionet (per exemple, *en-ca*) i la llengua del fitxer (ca). En la figura 9, per tant, «OpenSubtitles2018.en-ca.en» fa referència al fitxer monolingüe en anglès del subcorpus OpenSubtitles2018 en la combinació anglès-català. En cas que el paquet descarregat contingui un tercer fitxer amb l'extensió .ids, se'n pot prescindir per als propòsits d'aquesta activitat docent.

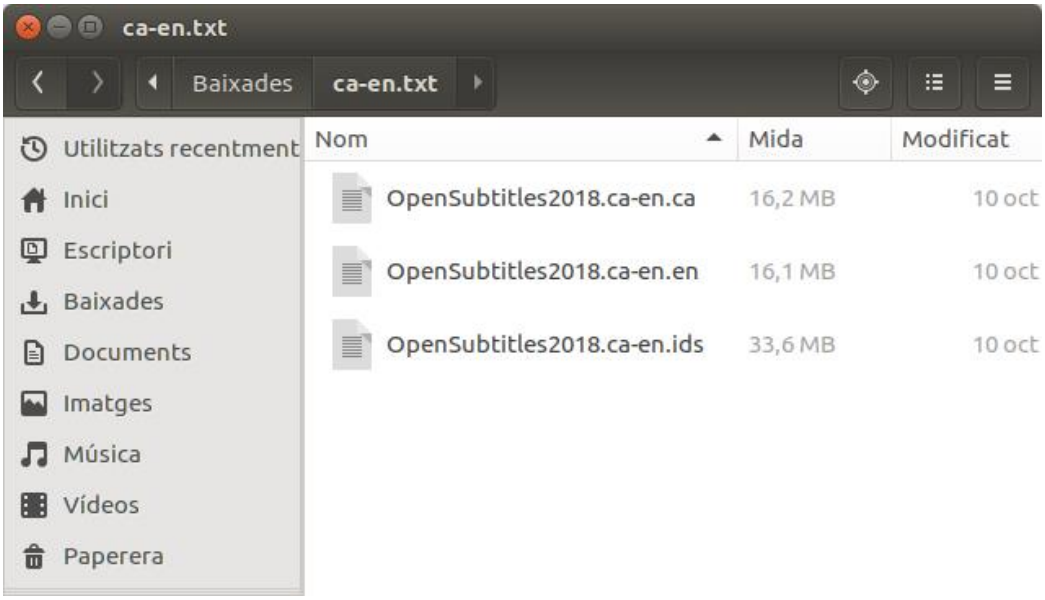


Figura 9: Contingut del paquet descarregat d'Opus.

Tots dos fitxers, editables amb qualsevol editor de text, mostraran la mateixa informació en L1 i en L2, en una frase per línia. Aquests són els dos fitxers que, sense necessitat de modificar-ne el contingut, caldrà carregar a la pestanya *Files* de MTradumàtica, mitjançant l'espai al peu de la pàgina. Un cop carregats, caldrà comprovar que MTradumàtica n'ha reconegut correctament les llengües, i l'usuari en podrà previsualitzar el contingut. La imatge següent mostra els botons que permeten dur a terme aquests dos processos:

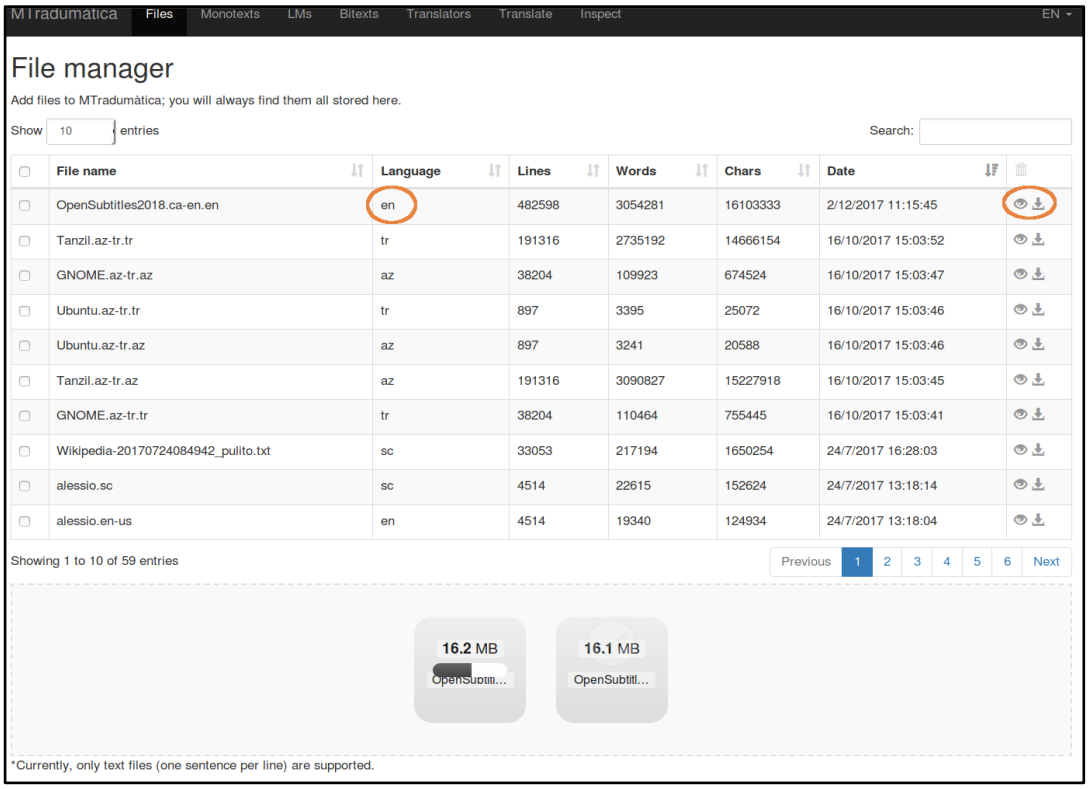


Figura 10: Càrrega de fitxers en MTradumàtica, modificació de la llengua del fitxer i previsualització del contingut.

4.2. Obtenció de corpus paral·lels en TMX i conversió a format Moses

L'activitat anterior es pot completar per mitjà de l'addició d'un altre corpus al model de traducció que es generarà en un pas successiu. Si els estudiants ja han tingut oportunitat de treballar en altres assignatures el format TMX, es pot aprofitar per remarcar que també disposen, per tant, de corpus paral·lels bilingües propis. En cas que no en tinguin cap, se'n poden trobar en diversos repositoris, com ara el de Softcatalà,[13] generat a partir dels projectes de traducció de productes lliures que l'associació ha anat localitzant de manera voluntària al llarg dels seus més de quinze anys d'existència. A diferència de la tasca anterior, en aquest cas caldrà convertir el corpus en format TMX al format apte per a Moses, la qual cosa es pot dur a terme fàcilment amb Okapi Rainbow.[14] En aquest programa, després d'haver importat (o arrossegat) el TMX a la pestanya Input List 1, caldrà indicar quines són les llengües que conté a la pestanya Languages and Encodings:

[13] Softcatalà <https://www.softcatala.org/recursos/memories.html/>
[14] Okapi Rainbow <http://okapiframework.org/>

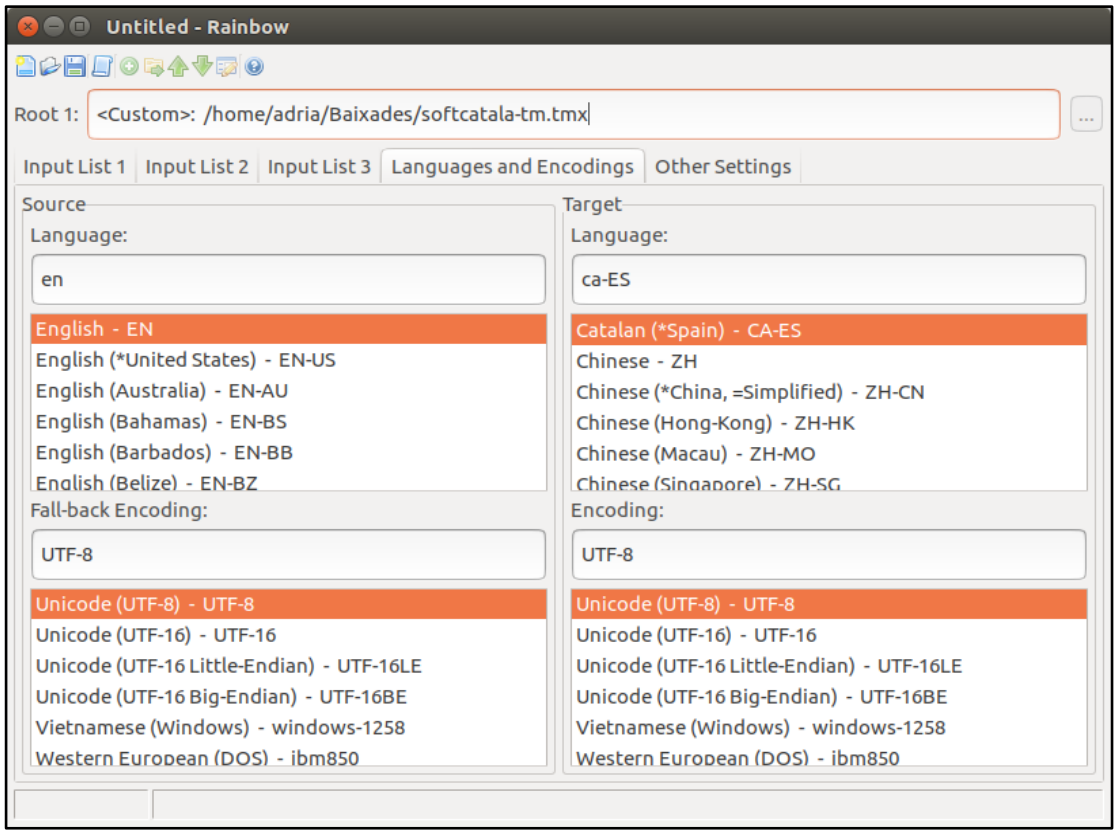


Figura 11: Pestanya Languages and Encodings d'Okapi Rainbow.

És important que els codis de les llengües seleccionats a Rainbow coincideixin exactament amb els codis que apareixen a la memòria, per a la qual cosa caldrà obrir el TMX amb un editor:

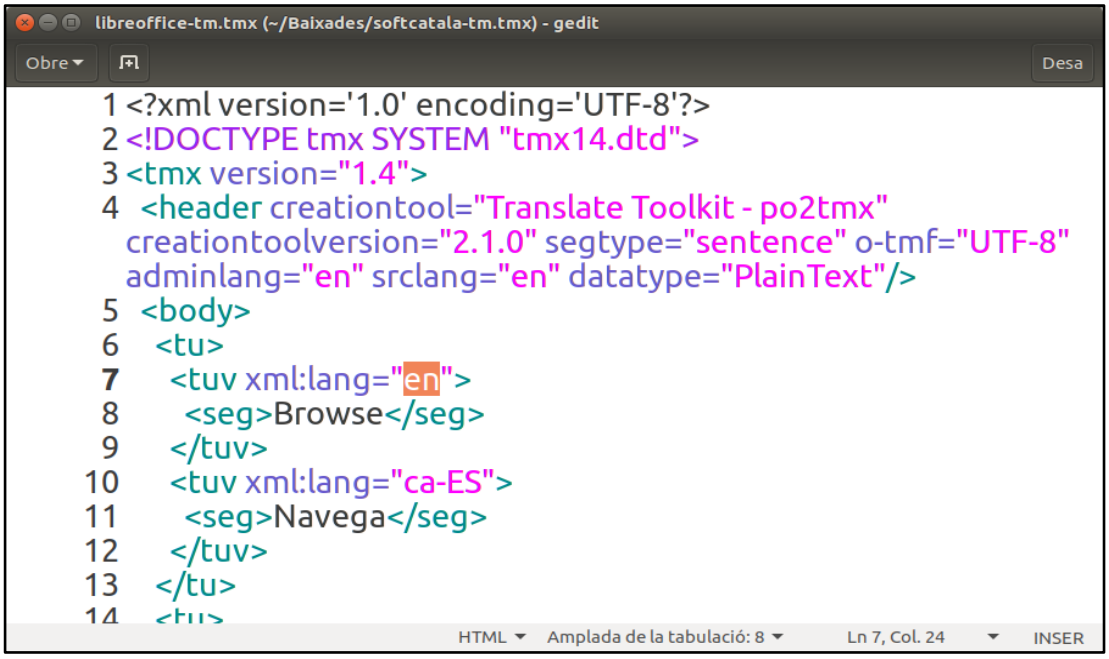


Figura 12: Codificació de llengües en un fitxer TMX. La primera línia també mostra la codificació (UTF-8 en aquest cas).

Un cop establertes les llengües i les codificacions en Rainbow, es pot fer servir la funció File Format Conversion (menú Utilities> Conversion Utilities) per convertir el TMX a format Parallel Corpus Files. El resultat de la conversió, com en la tasca anterior, seran dos fitxers de text amb extensions que corresponen a les llengües.

Figura 13: Resultat de la conversió d'un TMX amb Rainbow.

Com l'anterior, aquesta tasca acabarà amb la càrrega dels dos fitxers a MTradumàtica, la comprovació del reconeixement automàtic de les llengües i la previsualització del contingut.

4.3. Obtenció i neteja de corpus monolingües

Els corpus monolingües, que MTradumàtica utilitzarà per a crear un model de llengua real, es poden obtenir de diverses maneres. Cal tenir en compte, com en les tasques anteriors, que MTradumàtica només accepta fitxers en txt amb una frase per línia. En aquesta tasca, recorrerem a la Viquipèdia com a recurs per a l'obtenció de corpus monolingües.[15] La pàgina Exporta de la Viquipèdia[16] permet exportar categories d'articles a format XML.[17] Un cop feta l'exportació, caldrà netejar el corpus de codi, atès que contindrà etiquetes XML (<>) que no corresponen a text real. El procés de neteja es pot adaptar al nivell dels estudiants: en el cas de corpus molt grans, es poden fer cerques i substitucions automatitzades, incloent-hi expressions regulars;[18] altrament, en textos més breus, es poden seleccionar els fragments no desitjats i esborrar-los d'un en un. Un cop netejat el corpus, caldrà desar-lo en format txt i seguir els mateixos passos que en les tasques anteriors per carregar el fitxer a

[15] Vegeu, en aquest mateix número de la revista *Tradumàtica*, Oliver, Vàzquez i Ubide (2017) per a un mètode similar d'exportació de corpus de la Viquipèdia (http://wiki.dbpedia.org/).
[16] Wiquipèdia <https://ca.wikipedia.org/wiki/Especial:Exporta/>
[17] La llista de categories es pot consultar a https://ca.wikipedia.org/wiki/Especial:Categories/.
[18] Vegeu Martín-Mor i Peña-Irles (en premsa) per a un procés automatitzat de neteja de corpus de la Viquipèdia.

MTradumàtica, comprovar el reconeixement automàtic de llengua de MTradumàtica i previsualitzar-ne el contingut.

4.4. Generació de models i entrenament del traductor

Un cop carregats els fitxers de les tasques anteriors (un corpus paral·lel alineat descarregat d'Opus, un corpus paral·lel alineat convertit des de TMX i un corpus monolingüe exportat de la Viquipèdia), caldrà generar els models de llengua i de traducció, i finalment entrenar el traductor.

Per a la generació del model de llengua (per tant, monolingüe en llengua d'arribada), caldrà que, des de la pestanya Monotexts es creï un nou model de llengua per mitjà del botó +. Un cop assignat el nom del monotext i la llengua mitjançant el quadre de diàleg emergent, s'hi podrà afegir (per mitjà del botó + del nou monotext) el corpus monolingüe (en aquest exemple, el que correspon a la tasca 4.3, 4.3. Obtenció i neteja de corpus monolingües).

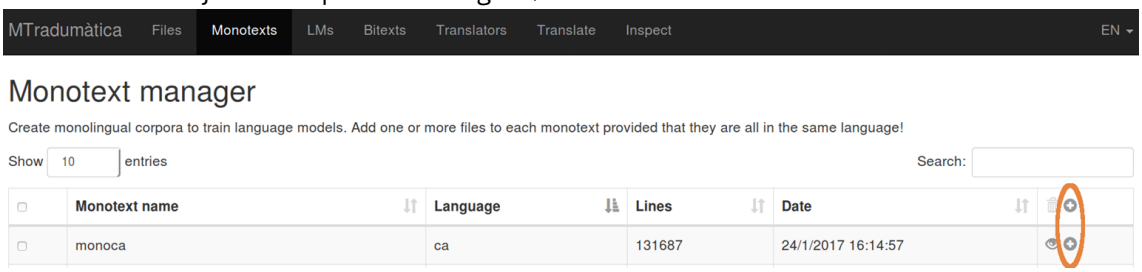


Figura 14: Creació de monotextos.

Prement un altre cop el botó + del nou monotext, es podran afegir tants fitxers monolingües (en la mateixa llengua) com es vulgui.

L'entrenament automàtic del model de llengua es durà a terme des de la pestanya següent, LMs, des d'on, mitjançant el mateix botó + de la fila de títol, es podran seleccionar els monotextos que es volen fer servir per al model

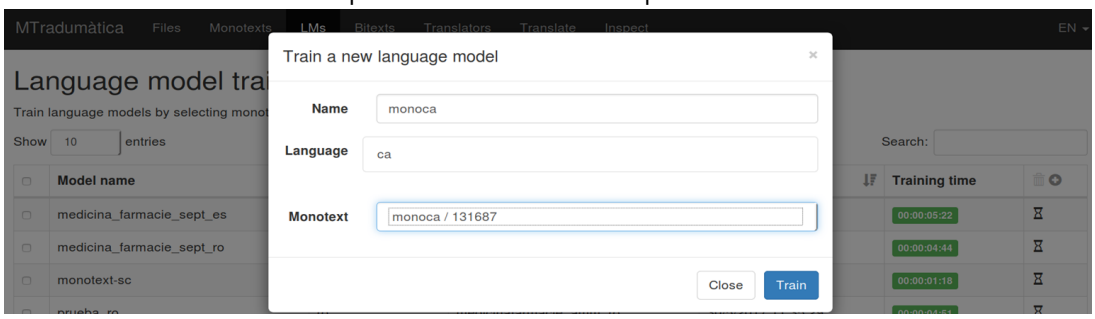


Figura 15: Entrenament de models de llengua.

El pas següent en l'activitat serà la creació de bitextos i la generació del model de traducció. El procés és el mateix que el que s'ha descrit en els paràgrafs anteriors per als monotextos: creació, des de la pestanya *Bitexts*, d'un bitext a partir dels fitxers corresponents a la tasca 4.1, 4.1. Obtenció de corpus paral·lels d'Opus i càrrega a MTradumàtica i combinació del bitext resultant amb els fitxers corresponents a la tasca 4.2, 4.2. Obtenció de corpus paral·lels en TMX i conversió a format Moses.

Finalment, a partir dels components anteriors (monotextos i bitextos), la pestanya *Translators* permetrà l'entrenament d'un motor de TA estadístic:

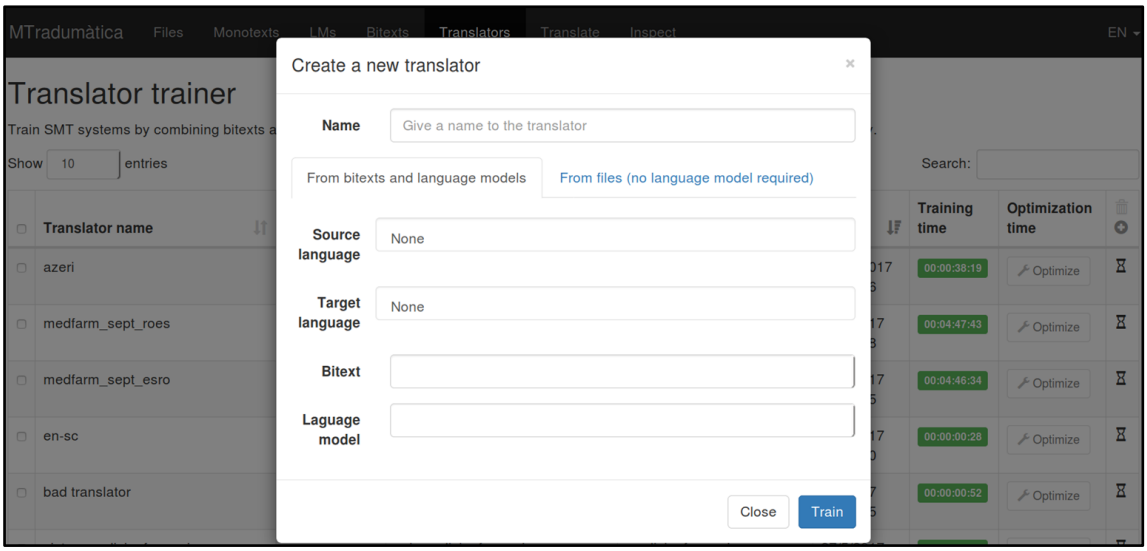


Figura 16: Entrenament de motors de TA.

Mitjançant el botó +, MTradumàtica obrirà un quadre de diàleg des d'on es poden triar el bitext i el model de llengua creats als passos anteriors.[19] En prémer el botó *Train*, començarà el procés d'entrenament (la durada del qual variarà en funció de la quantitat d'informació que continguin els fitxers carregats). Un cop finalitzat l'entrenament, la pestanya *Translate* mostrarà el traductor amb el nom que se li hagi assignat.

4.5. Activitat complementària — postedició de TA

L'activitat es pot complementar amb tasques relacionades amb la preedició i la postedició, sigui via web o via càrrega i descàrrega de fitxers, però també mitjançant la connexió de MTradumàtica a sistemes de SGET. Actualment, ja és possible connectar OmegaT a qualsevol instància de Moses que s'executi en un servidor, per la qual cosa també és possible tècnicament connectar OmegaT a MTradumàtica.

[19] El botó «From files (no language model required)», visible a la figura 16, permet generar un motor sense model de llengua. Més exactament, el que fa MTradumàtica és separar una part del fitxer del corpus bilingüe i reservar-la per generar el model de llengua.


OmegaT 4.1.3_1 :: prova
Projecte Edita Vés a Visualització Eines Opcions Ajuda
Editor - for_us...rings_en.strings
"Inglés" ;
"Switch to English";
<segment 2384>
"Contacts";
"Calls";
Coincidències parcials
1. "English";
"Inglés" ;
<50/33/50% >
Traducció automàtica
" Cambiar a inglés " ;
<Moses>
Diccionari
Diverses traduccions
Notes
Comentaris
Glossari
Segment Properties
El projecte s'ha desat automàticament a les 11:29
75% (236 left) / 86,5% (564 left), 5.433
20/0

Figura 17: Connexió d'un motor de MTradumàtica (Moses) a OmegaT.

El fet de dur a terme la PE des d'un sistema SGET permet al docent introduir conceptes com ara la PE automatizada —és a dir, la PE per mitjà de cerques i substitucions— o els diferents graus d'intervenció en el text durant la PE en funció de la visibilitat del text (PE de qualitat suficient per a l'assimilació de continguts; PE d'alta qualitat per a la disseminació de continguts), etc.[20]

5. Conclusions

El present article aborda la formació de traductors en l'àmbit de la traducció automàtica estadística apuntant fonamentalment la personalització i l'entrenament de motors basats en TAE. El punt de partida del treball han estat els resultats de l'informe de ProjecTA que entre d'altres aspectes identifica els punts sobre els quals cal actuar per a incrementar l'ús de la TA+PE, un dels quals, el de la formació. La formació en aquest àmbit requereix del coneixement del funcionament dels sistemes TAE integrats en un procés de treball més ampli de tal manera que els coneixements del posteditor integrin la gestió del procés.

El llindar de la formació en l'àmbit apuntat requereix d'eines que simplifiquin parts del procés. MTradumàtica, la plataforma descrita en l'article, té per objectiu proporcionar una plataforma web que integri fàcilment la personalització de motors de TAE en un procés de treball professional. Clarament, la plataforma ha de continuar evolucionant per mitjà de la integració de diverses funcions que en millorin la usabilitat (integració amb eines SGET, gestió d'usuaris, eines de postedició automàtica, gestió de terminologia, etc.).

[20] Vegeu Martín-Mor, Piqué i Sánchez-Gijón (2016: 69-70) per a les modalitats de PE; i Forcada (2009) per a les finalitats de la TA.

Mitjançant la proposta docent, els autors miren de col·laborar a un increment de la presència de la TA —i, específicament, de la personalització de motors de TAE— en els programes de formació de traductors. La proposta esmentada ha estat presentada de manera que pugui ser subdividida en mòduls i adaptada a diferents formats de classe. Per exemple, en la formació de grau, es pot utilitzar MTradumàtica per reforçar el coneixement declaratiu dels alumnes sobre la TA i il·lustrar conceptes relacionats amb la TA. En assignatures de traducció especialitzada o de tecnologies de grau i, especialment en la formació de màster, té sentit que els alumnes puguin dur a terme el procés d'entrenament d'un motor de manera parcial o completa. Per últim, en el pla dels programes de doctorat, MTradumàtica pot servir com a àrea de proves per a la recerca en TA.

En qualsevol cas, la inclusió de la TA en els programes de formació de traductors és tan sols una de les vies per a incrementar l'ús de la TA entre els traductors i, tal com indicaven els resultats de ProjecTA, les empreses i els professionals de la traducció necessitaran eines (i formació) personalitzables.

Bibliografia

Aziz, Wilker; Castilho, Sheila; Specia, Lucia (2012). PET: a Tool for Post-editing and Assessing Machine Translation. A *LREC* (p. 3982–3987). Recuperat de http://wilkeraziz.github.io/dcs-site/publications/2012/AZIZ+LREC2012.pdf [última visita: 14 de desembre del 2017].

Forcada, Mikel L. (2009). Apertium: traducció automàtica de codi obert per a les llengües romàniques. *Linguamática*, 1(1):13-23. Recuperat de http://www.linguamatica.com/index.php/linguamatica/article/view/18 [última visita: 14 de desembre del 2017].

Koehn, Philipp (2016). *Moses User Manual and Code Guide*. http://www.statmt.org/moses/manual/manual.pdf [última visita: 14 de desembre del 2017].

LT-Innovate (2013). *LT2013. Status and Potential of the European Language Technology Markets*. http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=4267 [última visita: 14 de desembre del 2017].

Machado, Maria José; Leal Fontes, Hilário (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/ [última visita: 14 de desembre del 2017].

Martín-Mor, Adrià (2017). MTradumàtica: Statistical machine translation customisation for translators. *Skase Journal of Translation and Intepretation, 11* (1), 25-40. Recuperat de http://www.skase.sk/Volumes/JTI12/pdf_doc/02.pdf [última visita: 14 de desembre del 2017].

Martín-Mor, Adrià i Peña-Irles, Víctor (en premsa). Creació d'un motor de TAE especialitzat en farmàcia i medicina per a la combinació romanés-castellà. *Linguamática 10*.

Martín-Mor, Adrià; Piqué, Ramon; Sánchez-Gijón, Pilar (2016). *Tradumàtica. Tecnologies de la traducció*. Vic: Eumo.

Oliver, Antoni; Vàzquez, Mercè; Ubide, Georgina (2017). Estudi de la fiabilitat de la Viquipèdia com a recurs terminològic. *Tradumàtica 15*. Recuperat de https://doi.org/10.5565/rev/tradumatica.193 [última visita: 14 de desembre del 2017].

Papineni, Kishore; Roukos, Salim; Ward, Todd, i Zhu, Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. A *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318). Association for Computational Linguistics. Recuperat de http://dl.acm.org/citation.cfm?id=1073135 [última visita: 14 de desembre del 2017].

Piqué Huerta, Ramon; Colominas, Carme (2013). Les tecnologies de la traducció en la formació de grau de traductors i intèrprets [en línia]. *Revista Tradumàtica 11*, 297-312. Recuperat de https://doi.org/10.5565/rev/tradumatica.43 [última visita: 14 de desembre del 2017].

Sánchez-Gijón, Pilar (2016). La posedición: hacia una definición competencial del perfil y una descripción multidimensional del fenómeno, *Sendebar 27*, 151-162.

Tiedemann, Jörg (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. A Nicolov, N., K. Bontcheva, G. Angelova i R. Mitkov (ed.), *Recent Advances in Natural Language Processing* (vol V). Amsterdam/Philadelphia: John Benjamins, 237-248. Recuperat de http://stp.lingfil.uu.se/~joerg/published/ranlp-V.pdf [última visita: 14 de desembre del 2017].

Torres-Hostench, Olga; Presas, Marisa i Cid-Leal, Pilar (coords.) (2016). *L'ús de traducció automàtica i postedició en les empreses de serveis lingüístics de l'Estat espanyol: Informe de recerca ProjecTA 2015*. Bellaterra. https://ddd.uab.cat/record/166753 [última visita: 14 de desembre del 2017]