Universitat Autònoma de Barcelona

Entrenament, avaluació i integració d'un traductor automàtic neuronal de l'àmbit biomèdic a una empresa

Marta Méndez Arroyo

Treball de Fi de Màster

Director: Antoni Oliver

Màster en Tradumàtica: Tecnologies de la traducció

Facultat de Traducció i Interpretació

Universitat Autònoma de Barcelona, 2019-2020

“Translation is that which transforms everything

so that nothing changes.”

GÜNTER GRASS

Resum

La traducció automàtica neuronal està, actualment, a l'ordre del dia de totes les agències del sector. En aquest treball de fi de màster es duu a terme la creació d'un motor de traducció automàtica neuronal, amb la combinació d'idiomes EN>ES, en l'àmbit biomèdic en col·laboració amb una agència de traducció i mitjançant la configuració de Marian. Es presenta tot el procés que comporta la creació d'un motor, inclosos la decisió de les llengües i l'àmbit, la cerca i preparació dels corpus, l'entrenament i l'avaluació. L'objectiu del treball és entrenar un motor de traducció automàtica que permeti augmentar la productivitat d'una empresa de traducció o d'un sector d'aquesta empresa i comprovar si ofereix millors o pitjors resultats que alguns dels traductors més potents del mercat, com són Google Translate i DeepL.

Paraules clau: traducció automàtica, traducció automàtica neuronal, entrenament, avaluació, textos biomèdics, agència de traducció, anglès, espanyol.

Abstract

Neural Machine translation is the new state-of-art in all the translation industry. In this Master's Degree dissertation, we create a neural machine translation engine with the language combination EN>ES about biomedicine in collaboration with a translation agency, built from Marian's configuration. We present all the processes that involves the creation of an engine, included the language and the field decision, the corpus research and preprocessing, the training and the evaluation. The aim of this dissertation is to train a machine translation engine that allows increase the productivity of a translation agency or part of it, and verify if it offers better or worse results than some of the best translation engines in the market, such as Google Translate and DeepL.

Keywords: machine translation, neuronal machine translation, training, evaluation, biomedical texts, translation agency, English, Spanish.

Resumen

La traducción automática neuronal está, actualmente, a la orden del día de todas las agencias del sector. En este trabajo de fin de máster se lleva a cabo la creación de un motor de traducción automática neuronal con la combinación de idiomas EN>ES, del ámbito biomédico en colaboración con una agencia de traducción y a partir de la configuración de Marian. Se presenta todo el proceso que implica la creación de un motor, que incluyen la decisión de las lenguas y el ámbito, la búsqueda y preparación de los corpus, el entrenamiento y la evaluación. El objetivo de este trabajo es entrenar un motor de traducción automática que permita augmentar la productividad de una empresa de traducción o de un sector de esa empresa i comprobar si ofrece mejores o peores resultados que algunos de los traductores más potentes del mercado, como son Google Translate i DeepL.

Palabras clave: traducción automática, traducción automática neuronal, entrenamiento, evaluación, textos biomédicos, agencia de traducción, inglés, español.

Índex de continguts

Índex de figures

Índex de taules

Agraïments

El present treball de fi de màster no hagués estat possible sense la col·laboració de moltes persones.

En primer lloc, vull agrair-li al meu tutor del treball, l'Antoni Oliver, l'esforç i la presència en cada moment, els recursos i les facilitats i, sobretot, la paciència.

A iDisc Information Technologies Inc., per l'oportunitat i la confiança que m'han brindat i per l'ajuda constant.

Als meus companys i companyes de professió, que han col·laborat sense pensar-s'ho en tot allò que els he demanat.

I finalment, a la meva família i els meus amics i amigues, que m'acompanyen cada dia i em recolzen en totes les meves decisions.

Sense vosaltres res d'això no hagués estat real.

1. Introducció i justificació

En aquest treball de final de màster es duu a terme l'entrenament d'un traductor automàtic neuronal de l'àmbit biomèdic per tal de poder-lo incorporar al flux de treball d'una empresa. Les tasques, però, no se centraran només en l'entrenament, sinó també en tots els processos que podem trobar abans i després, com són la recerca prèvia, la recopilació i el preprocessament de corpus, l'entrenament, les proves i les avaluacions. Com ja hem esmentat, es tracta d'incorporar aquest traductor automàtic al flux de treball d'una empresa, la qual cosa implica la presa de decisions abans de determinar quin serà el traductor automàtic que decidirem entrenar. El treball s'ha fet amb col·laboració d'una empresa de serveis lingüístics.

Partim d'una base en què som plenament conscients que la traducció automàtica està a l'ordre del dia a les empreses de traducció i serveis lingüístics, per tant, amb aquest treball acadèmic, pretenem confirmar que la productivitat augmenta gràcies a aquestes tecnologies, així com també demostrar la capacitat dels traductors automàtics que analitzarem.

Pel que fa al marc teòric del treball, farem una breu repassada pels diferents tipus de traducció automàtica que ens podem trobar actualment, atès que hem considerat que es tracta d'uns coneixements necessaris abans d'entrar a la part més tècnica que impliquen aquestes eines. També ens centrarem en les possibilitats d'avaluació dels sistemes de traducció automàtica i les mètriques que existeixen actualment i que es troben en major ús. Finalment, farem una breu explicació de la teoria que hem abordat sobre la introducció de tecnologies, en aquest cas traducció automàtica, dins del flux de treball d'una empresa, i les implicacions que això comporta.

Al pròxim apartat, presentarem els objectius i les hipòtesis que ens hem plantejat per aquest treball, els quals es relacionen amb els diversos àmbits que hem comentat, però sobretot en l'assumpte que ens ocupa, és a dir, l'acció d'entrenar i incorporar un traductor automàtic a una empresa del sector.

A continuació, mostrarem el marc teòric que ja hem comentat i posteriorment, a la metodologia, explicarem tot el procés que hem anat seguint, des de la primera recerca sobre motors de traducció automàtica fins a l'avaluació del motor entrenat. Es tracta d'un

procés llarg que hem dividit en set fases principals: la recerca sobre traductors automàtics del mercat, l'acotament de preferències, la decisió de l'àmbit i les llengües d'entrenament, la recopilació de corpus, el preprocessament de corpus, l'entrenament i l'avaluació.

Finalment, presentarem un breu apartat amb els resultats que hem obtingut i les conclusions extretes a partir de l'elaboració del treball i de tots el procés que s'ha anat seguint.

2. Objectius: preguntes de recerca i hipòtesi

A continuació, abans d'entrar en la matèria que ens ocupa, considerem necessari plantejar uns objectius pel treball, per així donar forma a aquelles preguntes que ens han sorgit i a les que volem donar resposta i també plantejarem una hipòtesi a partir de la qual es desenvoluparà el treball.

Les preguntes que ens hem plantejat són les següents:

- L'entrenament d'un motor de traducció automàtica d'un àmbit concret d'especialitat millora la productivitat d'una empresa?
- Val la pena invertir el temps i recursos en l'entrenament d'un motor de traducció automàtica completament personalitzat?
- Un motor de traducció automàtica completament personalitzat i entrenat amb corpus interns dona millors resultats que un altre motor genèric del mercat?

A partir d'aquestes preguntes i d'una petita recerca prèvia, la hipòtesi que presentem és la següent:

- A partir de l'entrenament d'un motor de traducció automàtica amb els corpus d'una empresa, la qualitat de la traducció automàtica serà millor que la d'un motor genèric i la productivitat de l'empresa augmentarà.

Així doncs, un cop plantejats els aspectes en què ens volem centrar, procedirem a endinsar-nos-hi i a aprofundir en les nocions teòriques que envolten aquest treball.

3. Marc teòric

A l'hora d'abordar les nocions teòriques d'aquest treball, obrim diverses línies de coneixement, ja que considerem que per tal d'abordar-lo necessitem entendre els diferents conceptes que envolten aquest àmbit. Començarem per plasmar una breu definició de la traducció automàtica que s'ha anat recuperant a partir de la bibliografia. A continuació, exposem una diferenciació entre els diversos tipus de traducció automàtica i seguirem amb teoria sobre l'avaluació de la qualitat de la traducció automàtica. Finalment, tancarem el marc teòric amb una petita aportació sobre la visió que hi ha de la introducció de la traducció automàtica a les empreses.

3.1. La traducció automàtica

A l'hora d'abordar un treball sobre la traducció automàtica (TA) hem de partir de la definició del concepte. Hi ha una gran quantitat de definicions en l'àmbit de les tecnologies i la tradumàtica, però ens centrarem en algunes d'elles. L'evolució de les noves tecnologies no només ha afectat la nostra vida quotidiana, sinó també la nostra professió: la traducció. S'han anat creant softwares que són capaços de traduir sols amb una productivitat, en teoria, major a la d'un traductor humà. Per tant, seguint la definició de Hutchins & Somers (1992:3), definim la traducció automàtica (el terme en anglès de la qual és *Machine Translation* (MT)), com a sistemes computacionals responsables de la producció de traducció d'una llengua natural a una altra, amb o sense assistència humana.

No només va ser el terme *Machine Translation* el que va aparèixer, sinó que també ens vam trobar amb altres com *mechanical translation* o *automàtic translation*, però actualment el que s'ha mantingut és el primer, així com també en espanyol o català, en què s'han unificat els termes *Traducción Automática* (TA) o *Traducció Automàtica* (TA) respectivament.

La principal idea de la traducció automàtica és la de produir traduccions d'una qualitat elevada, però, com esmenta Hutchins & Somers (1992:3), a la pràctica, el resultat de la traducció dels motors de TA es revisa (postedita) per un traductor humà. Dins de l'àmbit de la postedició hi ha diversos nivells, els quals depenen de les circumstàncies en què s'hagi fet ús de la TA i del propòsit del text.

Amb l'evolució de la traducció automàtica han anat apareixent diversos tipus de TA, i les preferències del mercat han anat variant d'uns a altres al llarg dels anys.

Per tal de posar-nos en context i comprendre com ha anat evolucionant la traducció automàtica al llarg del temps, presentem a continuació els diversos tipus de sistemes de traducció automàtica. De tipus n'hi ha molts, ja que cadascú ha elaborat la seva pròpia diferenciació i al llarg dels anys n'han anat apareixent diferents tipus, però nosaltres ens centrarem l'explicació de cinc tipus diferents; els tres primers presentats de manera breu i els dos últims de manera més extensa, ja que són els que ens interessen pel desenvolupament d'aquest treball. Per a fer aquesta diferenciació, ens basem en les explicacions d'Oliver (2014).

Diferenciem els següents tipus de sistemes de traducció automàtica:

- **Traducció directa**: es realitza la traducció a partir de la consulta de diccionaris bilingües i unes quantes regles per fer alguns canvis.
- **Sistemes de transferència:** es duu a terme una anàlisi del text de partida d'on s'extreu una estructura que, juntament amb unes quantes regles, es converteix en una estructura del text d'arribada.
- **Traducció *interlingua***: es fa una anàlisi del text en la llengua d'arribada a partir del qual s'obté una representació independent a la llengua. A partir d'aquí, es genera l'oració.
- **Traducció automàtica estadística**: basada en l'ús de corpus bilingües, que ajuden al sistema a "aprendre" a traduir.
- **Traducció automàtica neuronal**: es basa en la imitació de les xarxes neuronals humanes. A partir d'un entrenament amb corpus bilingües i una sèrie d'algoritmes, s'arriba a una traducció del text.

Per a l'explicació dels diferents tipus de sistemes, es farà ús d'un llenguatge planer i accessible, evitant entrar en els càlculs matemàtics que requereix l'explicació d'alguns dels sistemes. La intenció d'aquesta classificació és fer entenedora la diferència entre els diversos tipus, però evitant en tot moment fer referència a aquest coneixement matemàtic.

3.1.1. Traducció directa

Com el seu nom ens indica, amb aquest tipus de sistema es tradueix de manera directa, és a dir, traduïm a partir de la consulta de diccionaris bilingües, exclusivament, i l'ús d'unes quantes regles que permeten arribar a un text gramaticalment correcte. Podem dividir el funcionament d'aquests sistemes en dues parts, el procés d'anàlisi i lematització i el procés de generació de la traducció. A la primera part, es realitza l'anàlisi morfològica de l'oració en la llengua original i, a continuació, es duu a terme una lematització, és a dir, les diverses paraules es relacionen amb el seu lema, la qual cosa facilita cercar els mots al diccionari bilingüe i, per conseqüència, trobar-ne la traducció correcta. Un cop ha acabat el procés de lematització, es consulta el diccionari bilingüe i se n'extreuen les traduccions i, finalment, s'apliquen una sèrie de regles, com poden ser el canvi d'ordre dels mots, la concordança morfològica, la resolució de possibles ambigüitats, etc. i s'aconsegueix la traducció final.

Els primers sistemes de traducció automàtica que van aparèixer es basaven en aquest funcionament i, actualment, encara hi ha alguns sistemes al mercat que el mantenen, com per exemple Systran, que actualment està desenvolupant també la traducció automàtica neuronal, i OpenLogos.

A continuació, oferim una explicació simplificada del que seria l'anàlisi i la transferència de la llengua en aquest sistema. L'exemple està inspirat en el que se'ns ofereix a Oliver (2014):

Partim de l'oració catalana "La nena dibuixa una casa vermella", la qual volem traduir.

El primer pas és oferir una anàlisi morfològica i la lematització, que podria ser la següent:

La *La*	**nena** *nena*	**dibuixa** *dibuixar*	**una** *una*	**casa** *casa*	**vermella** *vermella*	**.** *.*
DA0FS0	NCFS000	VIMP3S0	DI0FS0	NCFS000	AQ0CS0	Fp

Taula 1. Anàlisi morfològica i lematització

A partir dels lemes obtinguts, es consultaria un diccionari bilingüe per tal d'obtenir la traducció de cada una de les paraules de l'oració:

la	nena	dibuixa	una	casa	vermella	.
the	girl	draw	a	house	red	.

Taula 2. Transferència lèxica

Un cop tenim la traducció, amb les etiquetes morfològiques que hem obtingut de l'anàlisi, es farien els canvis necessaris. En aquest cas, hauríem de modificar la forma del verb *draw* per la corresponent forma de la tercera persona del singular del present, que seria *draws*. A part de la concordança verbal, també hauríem d'aplicar la norma següent:

NC* AQ* → AQ* NC*

Aquesta norma ens indica que en anglès hi ha un canvi d'ordre en el sintagma format per un nom comú i un adjectiu qualificatiu i, per tant, hauríem de canviar l'ordre de *house red* per *red house.*

Per tant, l'oració final que obtindríem seria:

"*The girl draws a red house.*"

3.1.2. Sistemes de transferència

Els sistemes de transferència, també anomenats sistemes de transferència sintàctica, es basen en convertir l'estructura del text de partida en la del text d'arribada. El seu funcionament comença amb l'anàlisi sintàctica del text de partida, del qual se n'extreu una estructura que, a partir de diverses regles, es converteix en l'estructura del text d'arribada. Amb aquesta última estructura que es genera, es crea l'oració final traduïda. Podem dividir aquest procés en quatre parts:

- A la primera part, realitzem aquesta anàlisi sintàctica del text de partida, que tant pot ser un text com una oració.

- Seguidament, es fa una representació sintàctica amb regles de la llengua d'arribada, és a dir, es representa el text respectant la llengua original però construint ja l'estructura de la llengua d'arribada.

- A continuació es realitza el procés de transferència lèxica, on a partir de la consulta de diccionaris bilingües, obtenim la traducció dels mots.

- Finalment, té lloc la generació morfològica, és a dir, es realitzen els canvis morfològics necessaris per a respectar les regles de morfologia de la llengua d'arribada.

El fet de basar-se en l'anàlisi sintàctica de les oracions, fa que anomenem aquest sistema *sistema de transferència sintàctica*, atès que se centra en plasmar la sintaxis d'una llengua a una altra i resta importància a la transferència morfològica.

A continuació, mostrem un exemple simplificat basat, altra vegada, en el que se'ns mostra a Oliver (2014), on podrem veure les diverses fases del procés d'aquest sistema que hem comentat.

Continuem amb la mateixa oració que hem fer servir a l'apartat anterior: *La nena dibuixa una casa vermella*. Novament, busquem traduir aquesta oració a l'anglès.

Comencem amb la primera fase que es basa en l'anàlisi sintàctica de l'oració de partida:

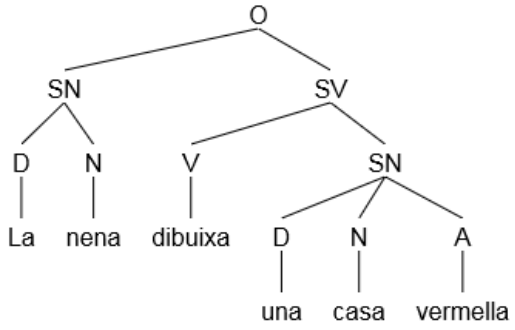


Figura 1. Anàlisi sintàctica de l'oració de partida

La segona fase es basava en una representació sintàctica aplicant les regles de la llengua d'arribada, com veiem a continuació, on comprovem el canvi d'ordre en el sintagma nominal que complementa al verb:

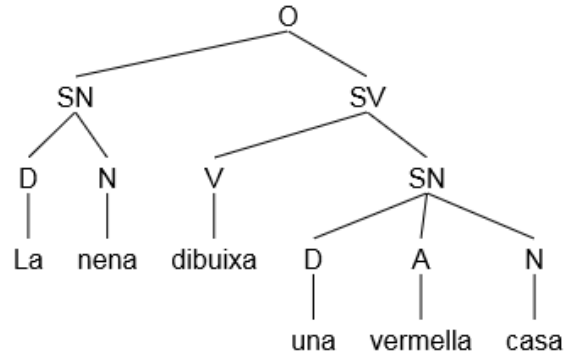


Figura 2. Representació sintàctica amb les regles de la llengua d'arribada

Es duu a terme la transferència lèxica i el resultat és el següent:

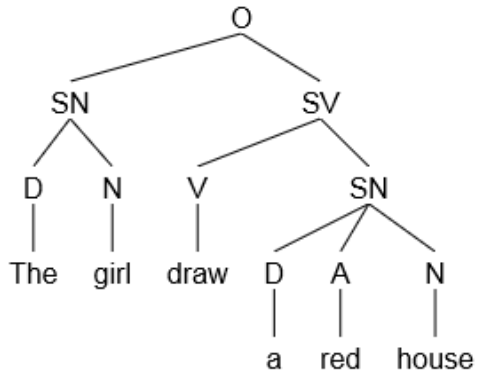


Figura 3. Arbre sintàctic amb transferència lèxica

Finalment, fem la generació morfològica on canviem la forma del verb *draw* per la tercera persona del singular del present (*draws*) i el resultat és:

"*The girl draws a red house.*"

3.1.3. Traducció interlingua

Els sistemes de traducció *interlingua* són, amb diferència, els que s'allunyen més del coneixement lingüístic que ens ocupa. El seu funcionament es basa, principalment, en l'anàlisi de l'oració en la llengua original, per tal d'obtenir una representació independent a la llengua, és a dir, podem afirmar que es centra en representar el significat independentment de la llengua (Trujillo, 1999). Un cop s'ha analitzat el text de partida,

aquest text es transforma en una representació abstracta (*interlingua*), i a partir d'aquest procés, es crea l'oració final.

Com a avantatge principal, cal esmentar que, a diferència d'altres sistemes, com per exemple el de transferència, aquest necessita menys punts de preparació, ja que es basa en un mòdul de sistema d'anàlisi i un de generació per cada llengua que treballem. Per tant, si volguéssim traduir de l'anglès al català i a l'espanyol, necessitaríem un mòdul de sistema d'anàlisi de l'anglès, un del català i un de l'espanyol i un sistema de generació de l'anglès, el català i l'espanyol, mentre que amb un sistema de transferència necessitaríem aquests mòduls i a més un sistema de transferència anglès-català i un anglès-espanyol.

La gran dificultat que presenta aquest sistema, però, és la de dissenyar aquesta representació intermèdia independent de la llengua, ja que és, com ja hem dit, abstracta.

A més, alguns d'aquests sistemes *interlingua* fan servir l'esperanto com a representació intermèdia, ja que es tracta d'una llengua d'estructura senzilla i rígida. No ens hem de confondre, per això, amb el terme *llengua pivot*, que seria la llengua que s'usaria entre la traducció d'una llengua A a una llengua B. L'ús de llengües pivot per a la traducció té avantatges i inconvenients; com a gran avantatge considerem el fet que requereix menys mòduls a preparar, però presenta un gran inconvenient i és que el fet de passar per dues llengües abans d'arribar a la d'arribada, augmenta la gran possibilitat de que apareguin més errors de transferència.

3.1.4. Traducció automàtica estadística (TAE)

A continuació, en aquest apartat i el següent explicarem els dos sistemes de traducció automàtica més usats actualment i, per tant, tractarem de fer-ho amb més detall que els anteriors. Tot i així, com ja hem esmentat anteriorment, procurarem no entrar en detalls matemàtics que podrien dificultar la comprensió de l'explicació.

La traducció automàtica estadística (*statistical machine translation* o SMT en anglès) es basa en l'ús de corpus paral·lels, és a dir, tradueix textos a partir d'informació extreta d'exemples d'altres traduccions que han produït traductors humans (Hearne and Way, 2011). Per tant, el sistema aprèn a traduir a partir d'uns models estadístics o paràmetres

que ha calculat a partir de textos i les seves traduccions. Aquesta tècnica es divideix en dos processos principals: l'entrenament (*training*) i la decodificació (*decoding*).

Seguint les explicacions de Hearne and Way (2011:208), a la fase d'entrenament hem de tenir en compte que per crear un TAE necessitem un model de llengua i un model de traducció. Pel que fa al model de llengua, es tracta d'una recopilació de corpus monolingües en la llengua d'arribada, que ofereixen al motor la informació sobre com es redacta en aquella llengua, cosa que permet crear traduccions més idiomàtiques o que semblin més pròpies de la llengua d'arribada. Per tot això, és important tenir un bon model de llengua, amb oracions del registre que ens interessa, que ens puguin ajudar a obtenir traduccions ben estructurades sintàcticament parlant. Pel que fa al model de traducció, es basa en la recopilació de corpus alineats en la llengua de partida i la d'arribada. Aquest últim s'utilitza per trobar totes les possibles traduccions d'un segment, per tant és important que sigui extens i contingui el màxim de traduccions possibles per fer més complet el motor.

Aquests sistemes no només combinen els dos models ja esmentats, sinó que també gaudeixen de la participació d'altres models, com són: els models de ponderació lèxica i reordenament, el nombre de paraules de traducció (*word penalty*) i el nombre de segments bilingües fets servir per traduir (*phrase penalty*). Pel que fa el model de ponderació lèxica, consisteix en l'ús d'un diccionari bilingüe probabilístic de paraules per determinar com de fiable és un segment bilingüe, mentre que el model de reordenament consisteix en condicionar el posicionament de les paraules en la llengua meta de la traducció dels segments utilitzats. Quant al *word penalty* i al *phrase penalty*, consisteixen en el recompte de paraules de la traducció i el recompte de segments bilingües usats per produir la traducció, respectivament. Tots aquests models es combinen, a cadascun d'ells se li assigna un pes diferents i a partir del càlcul dels pesos i la probabilitat, podem obtenir una puntuació final per valorar la nostra traducció.

Finalment, arriba la fase del *decoding* en què el TAE obté la oració en la llengua original, busca a l'espai totes les possibles traduccions, les puntua en base als diferents elements que hem esmentat anteriorment, calcula una puntuació final i tria la traducció amb la major puntuació, que és la que ens mostrarà.

La gran majoria de traductors automàtics estadístics que trobem al mercat estan basats en Moses (Koehn, P., Hoang, H., Birch, A., et al., 2007), que va aparèixer l'any 2005 sobre la base de models estadístics basats en conjunts de dades. Moses és un conjunt d'eines de codi obert, amb llicència LCPL, que ajuda a crear models propis i produir seqüències de dades. Donada la naturalesa de la nostra investigació, on el focus central és la traducció automàtica neuronal, no ens centrarem en el funcionament d'aquestes eines, però sí que es considera necessari el seu esment.

Així doncs, tot i que la traducció automàtica estadística encara està molt arrelada i present als sistemes d'avui en dia, l'aparició de la traducció automàtica neuronal és una realitat que s'ha de tenir en compte, ja que promet ser el futur de la tecnologia d'aquest sector. El seu funcionament el veiem a continuació.

3.1.5. Traducció automàtica neuronal (TAN)

Ja fa uns anys, va sorgir la traducció automàtica neuronal, i poc a poc ha anat agafant més força fins a convertir-se en el *new state-of-art*, és a dir, en la nova tendència en la TA. L'interès per la TAN, doncs, ha anat augmentant al llarg dels anys i, tot i que es basa en un procediment feixuc d'explicar, provarem d'explicar-lo des de la visió d'un traductor, evitant al màxim entrar en detalls molt tècnics. Cal tenir en compte que hi ha poques investigacions de caire acadèmic que aconsegueixin explicar el funcionament de la traducció automàtica neuronal sense entrar en termes tècnics, com per exemple Casacuberta i Peris (2017) i Forcada (2017). Forcada, doncs, aconsegueix explicar l'arquitectura neuronal de forma clara, per això ens basem principalment en les seves explicacions. Amb tot això, veiem que aquest alt nivell de dificultat dels articles, fa que el coneixement de la TAN s'allunyi dels professionals de la traducció que no tenen un perfil més tècnic i especialitzat.

La traducció automàtica neuronal (*neural machine translation* o NMT en anglès) parteix d'una base similar a la traducció automàtica basada en corpus o la TAE, ja que s'entrena amb grans corpus paral·lels, és a dir, amb grans memòries de traducció, però el funcionament és completament diferent; funciona a partir de xarxes neuronals (Forcada, 2017). El primer sistema de traducció automàtica neuronal va aparèixer l'any 2013, introduït per Nal Kalchbenner i Phil Blunsom.

Pel que fa a la representació de les paraules i les frases, fins ara els altres sistemes les representaven de manera discreta o abstracta (en el cas del sistema d'*interlingua*); la TAN les representa de manera numèrica mitjançant vectors (Casacuberta i Peris, 2017). Tot i basar-se en xarxes neuronals, el funcionament de la TAN té només una petita connexió amb el funcionament del cervell humà. Així doncs, potser s'hauria d'anomenar "traducció automàtica basada en xarxes neuronals artificials", ja que està composada de milers d'unitats artificials que imiten les neurones i l'activació que es realitza a partir d'estímuls que reben d'altres neurones; és a dir, aquestes neurones estan densament connectades entre elles i a partir d'una funció no lineal d'activació, la sortida d'una neurona és l'entrada de la següent. Es crea també, un vector de pesos que ajuda a calcular les possibilitats de traducció. Aquests pesos s'estimen a partir dels corpus d'entrenament que s'apliquen i una extensió d'un algoritme anomenat *algoritme de retro-propagació de l'error* (Casacuberta i Peris, 2017).

Les neurones es defineixen principalment pel seu comportament. Aquestes neurones tenen dues etapes de funcionament. Imaginem que volem realitzar l'activació X d'una neurona connectada a N neurones. A la primera etapa, cadascuna de les neurones es multiplica per un pes que representa la força i naturalesa de la seva connexió: w1, w2. w2, w4,...wN (Forcada, 2017). Aquests pesos poden ser positius o negatius, la qual cosa provoca que quan un estímul prové d'una connexió amb un pes positiu, la neurona activada tendeix a activar la neurona a què està connectada, i en cas que el pes sigui negatiu, la neurona activada tendeix a inhibir la neurona a la qual està connectada. D'aquí obtenim la següent equació, el resultat de la qual pot ser positiu o negatiu, però que encara no es tracta de l'activació de la neurona que ens ocupa:

$$y = w_1 \times x_1 + w_2 \times x_2 + w_3 \times x_3 + \ldots + w_N \times x_N + b,$$

A la segona etapa, s'apliquen les funcions d'activació, que poden ser de diferents tipus i, per tant, poden donar lloc a resultats molt diferents. Tot i així, cal tenir en compte que en moltes arquitectures de xarxes neuronals, les activacions de neurones no tenen sentit de manera individual, però un cop s'agrupen amb les activacions d'altres neurones, adquireixen aquest sentit.

A partir d'aquí, i gràcies a la incorporació d'aquests grans corpus d'entrenament, cada paraula es converteix en la representació d'ella mateixa i de tota la informació que

l'associa. El sistema relaciona aquesta informació que s'associa a cada paraula i la informació que impliquen les paraules d'una oració i, juntament amb els mecanismes de l'aprenentatge automàtic, aprèn a traduir. Així doncs, el sistema aprèn que dues paraules, a partir de la informació que les envolta, són termes més similars entre sí que d'altres que no tenen relació, cosa que li dona informació sobre les traduccions que ha de triar (Parra-Escartín, P., 2019).

Actualment, el model més utilitzat dins de la traducció automàtica neuronal és l'estructura codificador-decodificador amb atenció. Aquesta estructura permet que els models que es creen puguin seguir el procés d'entrenament, avaluació i traducció (Junczys-Dowmunt, Grundkiewicz et al., 2018), és a dir, processa la informació dels corpus bilingües que es relacionen a partir de l'aprenentatge automàtic, xarxes neuronals profundes i un sistema d'atenció agregada. Aquest sistema d'atenció analitza el context de cada paraula, com explicàvem al paràgraf anterior amb les indicacions de Parra-Escartín, P. (2019).

La principal diferència que trobem entre els sistemes de traducció automàtica estadística i els neuronals, és que l'element central dels primers són les paraules (o n-grames), mentre que l'element central dels segons són els vectors. A més, la gran avantatge de la traducció automàtica neuronal i d'aquests vectors és la facilitat d'aprenentatge del text gràcies a les xarxes neuronals recurrents, on una absorbeix el text d'entrada o l'altra produeix el text de sortida.

En el nostre cas, l'entorn que hem emprat per entrenar el nostre traductor automàtic, és el conegut Marian. Marian és un entorn de programació desenvolupat en C++, de codi obert i amb llicència MIT. És, de fet, la base de la majoria de traductors automàtics neuronals que hi ha actualment al mercat. Està desenvolupat a partir de la funció estadística del sistema Moses i el seu funcionament es basa en la xarxa codificador-decodificador comentada anteriorment. La xarxa end-to-end que utilitza contempla textos bilingües per maximitzar les probabilitats de traducció, com ja hem comentat anteriorment.

Per tal de poder seguir el procés, hem de tenir en compte alguns conceptes imprescindibles que formen part del procés de preparació dels corpus per poder entrenar un motor de traducció automàtica neuronal. Alguns d'aquests processos són necessaris tant per l'entrenament de sistemes neuronals com per estadístics. Procedim, doncs, a

enumerar-los i definir-los breument, ja que al marc metodològic podrem comprovar com són i per què es necessiten.

El primer concepte és la tokenització. El procés de tokenització es basa en la separació de les unitats lèxiques (Chung, T i Gileda, D, 2009). Es tracta del primer pas que hem de fer a l'hora d'entrenar un traductor automàtic estadístic o neuronal. Com comenten Cung i Gileda, hi ha llengües que no utilitzen espais entre paraules, com el xinès i, per tant, segmentar les oracions sol ser un procés difícil. Hi ha altres llengües, com la nostra, que sí que fan servir l'espai entre paraules, però dins d'aquestes paraules hi ha diferents morfemes, que a vegades corresponen a paraules independents. El procés de tokenització s'encarrega, en les nostres llengües principalment, de la separació entre les paraules i els signes de puntuació, però també es pot encarregar de separar paraules compostes en els seus diversos morfemes.

El segon concepte que ens ocupa és el procés de truecasing. El procés del truecasing consisteix en recuperar la informació sobre majúscules i minúscules del text brut (Lita, L.V., Ittycheriah, A., et al., 2003). Aquest fet millora la llegibilitat del text a l'hora d'entrenar, la precisió de les dades que aporten les majúscules i minúscules i normalitza el seu ús depenent de l'estil, la font i el gènere. És a dir, aquest procés assigna a cada token les majúscules i minúscules tal i com li corresponen, per tant, només s'assignaran majúscules a aquelles paraules que apareixen així al diccionari i als corpus que fem servir per entrenar, és a dir, les fonts i l'estil, independentment de la seva posició en l'oració.

Finalment, dos últims conceptes que hem de tenir en compte són el càlcul de subwords i l'aplicació de BPE (*byte-pair-encoding*). Com bé sabem, l'entrenament dels sistemes de traducció automàtica neuronal depèn, en gran part, del gran volum de vocabulari que conté. El problema sorgeix quan aquest vocabulari es basa en paraules fixes, la qual cosa vol dir que si no apareix la paraula al corpus d'entrenament, el sistema la classificarà com a "out-of-vocabulary", és a dir, fora del vocabulari (Kudo, T., 2018). Per a solucionar-ho, es separen les paraules en *subwords*, és a dir, en unitats més petites que la paraula i s'indexen per tal d'incorporar-les a l'entrenament. El model d'algorisme de segmentació de paraules estàndard és el BPE o *byte-pair-encoding* (Sennrich et al., 2016). El *Byte Pair Encoding* (BPE) (Gage, 1994) és un algorisme de compressió de dades simple, que adaptem a la segmentació de paraules, que reemplaça iterativament el parell de bytes més

freqüents en seqüències de caràcter amb un sol byte que no ha estat utilitzat (Sennrich et al., 2016).

A part d'aquests processos explicats, el preprocessament dels corpus per l'entrenament inclou altres passos, que s'expliquen més endavant al marc metodològic, com és la neteja de corpus, el tractament d'expressions numèriques i la divisió dels corpus en entrenament, validació i avaluació. Considerem que aquests aspectes no demanen una explicació teòrica, atès que és la pràctica el que els defineix.

3.2. Avaluació i mètriques

Pel que fa l'avaluació de la traducció automàtica, hi ha dues grans maneres de dur-la a terme: de manera automàtica o de manera manual. En el nostre cas ens centrarem en l'avaluació automàtica quantitativa, és a dir, que dona un valor numèric a la qualitat, i l'avaluació manual qualitativa, que detalla amb profunditat els errors que podem trobar a les traduccions.

Quant a l'avaluació automàtica quantitativa, al 2002 els tipus d'avaluació que hi havia eren cars i suposaven una inversió de temps massa gran (Pan Mae, H., 2016). Per això, els científics Kisshora Papineni, Salim Roukos, Todd Ward i Wei-Jing Zhu van presentar un nou mètode que van descriure com més ràpid, més econòmic i independent de la llengua. Aquest sistema s'anomenava Bilingual Evaluation Understudy (BLEU), que parteix de la màxima que com més a prop estigui el resultat de la TA de la traducció humana amb què es compari, més bo serà. Aquest sistema, partia de la base en que es comparaven els *n*-grames de la traducció automàtica amb els *n*-grames de la traducció de referència, independentment de la seva posició dins de l'oració, simplement, quantes més coincidències, millor puntuació. La investigació, però, va anar avançant i la posició dels *n*-grames i les paraules de què estaven rodejats van adquirir més importància dins del càlcul i li va atorgar una major precisió. A partir d'aquí, a través d'unes fórmules, es fa un càlcul de la qualitat de tot el corpus d'avaluació tenint en compte frase per frase. La puntuació és del 0 a l'1, sent l'1 la millor puntuació. Per a dur a terme aquesta avaluació, com es pot intuir, es necessita l'oració resultant de la traducció automàtica i un corpus de traduccions humanes de referència (Pepineri et al., 2002).

Malgrat ser la mètrica per excel·lència, BLEU té bastants inconvenients. El primer de tot i més important és que la llengua no és una ciència exacte, és a dir, hi ha moltes traduccions vàlides d'un sol text original, i ens podríem trobar amb el fet que BLEU penalitza una traducció perfecta només pel simple fet de ser diferent a la de referència.

Així doncs, existeixen altres mètriques d'avaluació molt importants, com són per exemple la NIST, WER i TER. Pel que fa a la mètrica NIST, obté el nom de les sigles de l'Institut Nacional de Normes i Tecnologies dels Estats Units (National Institute of Standards and Technology) i deriva del BLEU, tot i que té una diferència; tot i avaluar les coincidències de *n*-grames com BLEU, dona millor puntuació al sistema de traducció que és capaç de proporcionar una traducció correcta i en l'ordre correcte (Zhang et al., 2004). Així que tot i avaluar traducció automàtica, es centra en similituds i no en qualitat, igual que el BLEU.

Quant a la mètrica WER (Word Error Rate), medeix la quantitat mínima de paraules que s'eliminen, afegeixen o substitueixen per transformar una oració en una altra. El problema principal d'aquesta mètrica és que es considera la més simple de totes ja que, donat el fet que entre diverses traduccions hi ha diferències en l'ordre de les paraules i els complements sense que el significat canviï, la mètrica WER penalitza aquests canvis d'ordre amb la mateixa puntuació que si es tracta d'un error en una paraula, la qual cosa fa que el resultat sigui imprecís i poc fiable (Babych, B., 2014).

Pel que fa a la mètrica TER (Translation Edit Rate), medeix el nombre d'edicions que haurà de fer un traductor humà per canviar el resultat de la traducció automàtica i fer que aquesta coincideixi exactament amb la traducció de referència, és a dir fer que sigui el màxim de fluida i adequada al significat. El resultat d'aquesta mètrica s'obté a partir de la fórmula que veiem a continuació, on es divideixen el nombre de canvis realitzats pel nombre de paraules del text de referència. Entenem com a canvis o edicions qualsevol paraula o grups de paraules insertades, eliminades, substituïdes o canviades de posició dins l'oració (Snover et al., 2006). Observem, doncs, la fórmula esmentada:

$$\text{TER} = \frac{\text{\# of edits}}{\text{average \# of reference words}}$$

Il·lustració 1. Fórmula de càlcul de TER

Per tal que el resultat de la mètrica TER sigui favorable, ha d'acostar-se el màxim possible a 0, tenint en compte que totes les edicions esmentades anteriorment tindran el cost d'1.

Hem de tenir en compte que totes aquestes mètriques parteixen del text de referència com a model de traducció correcta, per tant, donat aquest fet, l'ús del vocabulari i l'estil similar al text de referència és l'única cosa que ens garantirà una qualificació alta (Tomás et al., 2003).

Ja sigui d'una manera o d'una altra, l'avaluació requereix tres elements bàsics: el text o oració original, el text o oració traduït pel traductor automàtic que volem avaluar i el text o oració traduït per un traductor humà, que servirà de referència.

Quant a l'avaluació manual o humana i qualitativa, partim de la base que ens ofereix Guzmán et al. (2015) que plasma els pensaments d'altres autors i junts afirmen que és necessària la participació d'humans per a avaluar els resultats de la traducció automàtica per dues raons; la primera és que les traduccions automàtiques estan destinades a una audiència humana, i la segona és que l'enteniment del món que tenim els éssers humans permet detectar els errors importants que fan els sistemes de TA. Per tal de dur a terme aquesta avaluació, podem fer servir diversos mètodes. A continuació en mostrem tres.

El primer sistema d'avaluació que podem dur a terme és el proposat per White et al. (1994), on s'avaluen dues coses diferents de la traducció: l'adequació i la fluïdesa. Pel que fa a l'adequació, els avaluadors o posteditors, etiqueten la capacitat del traductor de traduir el text, és a partir de segments de la traducció amb context que els permeti trobar-se en una situació "real", determinen si el TA ha traduït tot, part o res del text original. Els nivells que ho determinen solen ser d'entre 4 i 5, que parteixen de la traducció completa del text original fins a cap mena de traducció respecte l'original. Quant a la fluïdesa, els mateixos avaluadors determinen la llegibilitat del text traduït, és a dir, si les frases del text estan ben construïdes i permeten que el text es llegeixi de manera fluida, sense tenir en compte si és una traducció correcta o no.

El segon sistema d'avaluació humana que presentem és el que proposen Vilar et al. (2007). Aquest sistema es basa en la comparació de dos o tres sistemes de traducció automàtica i en la classificació del millor i el pitjor. La comparació es duu a terme a partir d'un conjunt de frases que s'han traduït amb dos o tres motors de TA diferents. Els

avaluadors han de seleccionar la millor traducció per a cada segment o marcar si les traduccions són igual de bones o de dolentes.

Finalment, el tercer sistema d'avaluació és el que explica Popovic, M. (2018), que parla de la classificació d'errors i l'anàlisi per a l'avaluació de sistemes de TA, proposat per Vilar et al. (2006) i anteriorment presentat Litjós et al. (2005). Explica que en aquest sistema d'avaluació, els avaluadors seleccionen els errors que troben a la traducció i els classifiquen en diverses categories prèviament estipulades. És el tipus d'avaluació que requereix més temps i recursos, i a vegades els resultats poden ser massa diferents i confusos, però permet comprovar en quins punts el motor de TA té més errors, per poder així millorar-lo o també poder plasmar aquests errors en una guia de postedició que faciliti la tasca als posteditors i millori la productivitat. Aquest sistema d'avaluació, com ja hem dit necessita una classificació concreta dels errors, que permeti guiar l'avaluació i facilitar-ne els resultats. En el nostre cas, ens centrarem en la classificació d'errors que presenta Panić, M., (2019), responsable del desenvolupament de la pàgina web de TAUS, basada en el seu sistema DQF-MQM, que l'expliquen de la següent manera:

> "It provides a comprehensive catalog of quality error types, with standardized names and definitions and a mechanism for applying them to generate quality scores." (Panić, M., 2019)

Com veiem, aquest model presenta la seva categorització d'errors i un número de gravetats que podem aplicar a aquests errors, que inclouen, *ciritcal, major, minor,* i *neutral* i van des del tipus d'error més greu a un tipus d'error basat en la preferència del posteditor, que no implica un error en si.

En el nostre cas, però, hem fet ús del model que fa servir l'empresa amb qui col·laborem, que és una adaptació del model explicat. Aquesta adaptació es mostra a l'Apartat 4.7.2.1. Categorització d'errors.

3.3. Incorporació de la TA al flux de treball d'una empresa

Al llarg dels anys, des que les tecnologies van començar a incorporar-se a la labor de la traducció, s'han dut a terme diversos estudis sobre l'afectació d'aquestes eines tecnològiques. En aquest apartat, però, ens centrarem en com s'ha estudiat la incorporació

de la TA a les empreses de serveis lingüístics. Partim d'un primer article d'Arevalillo (2012) on ens menciona la visió que hi havia llavors de la traducció automàtica a les empreses i, sobretot per part dels traductors. Des d'un bon inici els traductors consideraven que la TA estava dissenyada per destruir i substituir la figura del traductor, però poc a poc la societat es va conscienciant que la TA no és un enemic, sinó un aliat, que ens permet aconseguir encàrrecs i projectes que no es podrien haver fet abans, ja sigui pel cost, pel temps, etc. Així doncs, mantenim la idea que la traducció automàtica reforçarà la feina del traductor, però que aquest últim mai desapareixerà, ja que, com afirma Arevalillo (2012): "en la traducción entran en juego procesos mentales, gramaticales, semánticos y contextuales difícilmente interpretables 100% por una máquina, o al menos con una garantía mínima de fiabilidad".

Aquestes idees, igual que la tecnologia, avancen ràpid al llarg dels anys, i ens trobem que cada cop més empreses de serveis lingüístics han incorporat la TA al seu flux de treball. Han passat ja gairebé cinc anys des de l'elaboració de l'informe de recerca ProjecTA (2015), on ja veiem que el 45,5% de les empreses enquestades feien ús de la TA, encara que fos en poca mesura. De ben segur que si ara es tornés a fer el mateix estudi amb les mateixes empreses, la xifra augmentaria. En aquell moment, un 52,7% de les empreses enquestades no feien servir mai TA per diverses raons, com per exemple, la poca sol·licitud de TA per part dels clients, la incomoditat que suposava pels traductors, però, sobretot, la manca de confiança en la qualitat de la TA i la reducció de les tarifes que suposava aplicar-la.

Tot això ha anat canviant, i més a mesura que la tecnologia avança. Quan es va realitzar l'informe ProjecTA, només el 16% de les empreses enquestades tenien un sistema de TA propi, la qual cosa és ben segur que ha augmentat considerablement en aquest cinc anys i més amb la millora de la traducció neuronal. En aquell moment, la majoria d'empreses es decantaven pel sistema estadístic davant del sistema de traducció automàtica basada en regles, ja que els resultats eren considerablement millors. Actualment, però, la tendència es decanta cap a la traducció neuronal, que ha mostrat resultats increibles i que segur que encara millorarà considerablement.

Un últim aspecte que hem de mencionar és l'aparició de la figura del posteditor. Ja Arevalillo (2012), i segurament, altres autors, informaven sobre l'aparició d'aquesta figura i la seva ambigüitat. La figura del poseditor no és res més enllà de la figura del

traductor amb unes nocions tecnològiques que fins ara no li havia calgut tenir. Actualment, es comença a incorporar a la docència de traductors l'àmbit de postedició i, a més, moltes empreses formen als seus traductors en aquesta branca, per tal de donar-los més ventall de possibilitats. També és cert, que per les empreses que han decidit tenir un traductor automàtic propi és una gran oportunitat formar els seus traductors, ja que els poden formar amb els errors i matisos que necessiten controlar del seu traductor automàtic, la qual cosa augmentarà la productivitat a uns nivells molt elevats.

Així doncs, si bé és veritat que incorporar la TA al flux de treball d'una empresa és un procés lent, que pot comportar problemes al principi, amb una inversió en la capacitat tecnològica i humana, una empresa pot optar per entrenar el seu propi TA, incorporar-lo a l'empresa i veure com la seva productivitat augmenta progressivament.

4. Marc metodològic

Com ja s'ha esmentat anteriorment, aquest projecte es divideix en diverses fases, ja que es tracta d'un procés lent i elaborat que s'ha hagut de dividir en unes fases estrictament marcades per tal de seguir un ordre i obtenir uns resultats exitosos.

Aquestes fases, doncs, es podrien dividir de la següent manera: recerca sobre els diversos traductors automàtics del mercat, acotament de tria segons les preferències de l'empresa, decisió de l'àmbit d'interès i les llengües que es volien tractar, recopilació i neteja de corpus, entrenament i avaluació.

4.1. Recerca sobre traductors automàtics del mercat

Vam començar amb una cerca dels diversos traductors automàtics que hi ha al mercat, ja siguin privatius o lliures, de pagament o gratuïts, o tant si permeten personalització com si no.

Aquesta cerca es va dur a terme de manera extensa, i es va arribar a recopilar informació de més de trenta motors de traducció automàtica diferents, tot i que més endavant en vam seguir descobrint de nous.

4.1.1. Llistat de traductors automàtics del mercat

L'objectiu d'aquesta cerca era investigar per trobar el major número de solucions que el mercat ofereix actualment quant a traducció automàtica, sigui del tipus que sigui. En aquest primer punt de la investigació, simplement es va realitzar una llista de tots els noms de traductors automàtics que descobríssim, sense mirar-ne les característiques.

Aquesta llista es troba a l'Annex 1, on observem una taula amb tots els noms de traductors automàtics que es van trobar, encara que més endavant en van sorgir d'altres que ja no es van anotar.

4.1.2. Taula dels motors amb les característiques

A continuació, vam elaborar una taula amb tots aquells motors que havíem trobat, o la gran majoria d'ells, i totes les característiques que esmentarem a continuació, que són les que des de l'empresa es van considerar més rellevants per tenir en compte a l'hora de considerar per quins motors ens decantaríem. Les característiques són les següents:

a. Ubicació dels servidors del traductor automàtic: Era un dels aspectes a tenir més en compte, ja que els països de la UE han de complir amb la RGPD (Reglament General de Protecció de Dades) i molts dels clients de l'empresa requereixen un nivell de seguretat en les seves dades molt important, per tant, havíem de comprovar on s'ubiquen els servidors de cada traductor automàtic, i veure si compleixen amb la RGPD i on s'emmagatzemaven totes les dades.

b. Els volums necessaris per treballar amb el motor: observar quines recomanacions ens feien respecte el volum mínim de dades per entrenar el motor i quins tipus de corpus admetien. Tot això tenint en compte que permetessin l'entrenament, ja que molts dels motors que es van analitzar en un primer moment no disposen d'aquesta funció.

c. El cost del servei i com es quantifica: comprovar que hi ha serveis gratuïts, i veure quins són. Pel que fa els que són de pagament, veure com es quantifica aquest servei, ja sigui mensualment, com per paraula traduïda i també quin seria el cost de l'entrenament en aquells que ho permeten.

d. Altres serveis que s'oferissin al voltant de la traducció: Alguns dels motors de traducció automàtica estan integrats en solucions en el núvol i que ofereixen altres tipus de complements. Altres estan integrats dins de eines de traducció pròpies. Hi ha TA que també integren eines d'avaluació automàtica a partir de mètriques. Vam optar per llistar per cada motor els diferents serveis col·laterals que poden oferir, sense detallar-los, ja que això ho comprovaríem en cas que el motor fos seleccionat per l'entrenament. Vam considerar que era un punt important que podia afavorir l'elecció d'un TA.

e. Integració del TA a eines de traducció assistida per ordinador (TAO): No només observar si el TA es pot integrar a eines TAO, sinó també veure de quina manera

es duu a terme la traducció, si pujant memòries i obtenint-ne la traducció o si integrant-la a les eines de treball dels traductors.

f. Tipus de format d'arxius que admet: Observar quins formats admet i quins processos hauríem de fer per poder treballar amb aquesta eina. Per exemple, si només admet arxius monolingües, què retorna i com ho podem importar a les nostres eines? A més, observar també quins tipus de format d'arxius admet per l'entrenament del motor.

g. Les llengües: Veure les llengües que suporta cada TA. Per aquells TA que s'entrenen, veure si ens permet entrenar totes les combinacions de llengües que vulguem o si restringeix l'entrenament a certes llengües.

A l'Annex 2, podem veure una taula amb totes aquestes característiques esmentades per cada motor analitzat. Observem també, que hi ha aspectes que no s'han completat, ja que hi havia motors que no disposaven de documentació i, per tant, no ens informaven dels aspectes que volíem tenir en compte. Bastants dels motors esmentats a l'apartat anterior no apareixen a la taula per la ja comentada manca d'informació.

4.2. Acotament de preferències

En haver descrit les opcions que havíem anat trobant del mercat actual, va arribar el moment de prendre una decisió i veure per quina, o quines, opcions ens decantàvem. La primera presa de decisions va ser fàcil, ja que havíem descobert un traductor automàtic neuronal que es podia personalitzar, MTUOC, en el qual hi treballa el director d'aquest treball, així que es va considerar com a primera opció a explotar, ja que personalitzar un TA des de zero ens assegurava que el resultat podria ser completament a la nostra mida. Així doncs, vam decidir entrenar el traductor automàtic MTUOC a l'empresa.

Tots els processos que comentarem a continuació, es troben documentats en línia a la referència corresponent a Oliver, A. (2020a), i tant la preparació dels corpus, com l'entrenament i l'avaluació han anat seguint aquests passos.

4.3. Decisió de l'àmbit i les llengües

Després de veure les necessitats de l'empresa, vam observar que l'àmbit que més podria necessitar la incorporació de TA, per tal d'augmentar la productivitat, era l'àmbit de la traducció biomèdica. A més, vam acotar les llengües d'ús, i vam veure que actualment dins d'aquest camp, la combinació de llengües que més interessava a l'empresa per entrenar era la combinació anglès<>espanyol (EN<>ES), en ambdues direccions, tot i que a les explicacions ens centrarem en l'entrenament EN>ES. A partir d'aquí, doncs, va començar la següent fase, en què recopilaríem diversos corpus d'aquest àmbit d'especialitat i d'aquesta combinació de llengües.

4.4. Recopilació de corpus

La fase de la recopilació de corpus es va dividir en dues vessants: la recopilació de corpus interns de l'empresa i la de corpus públics externs. De la primera vessant se'n va encarregar el departament tècnic de l'empresa, que es va dedicar a recollir totes les memòries de traducció de projectes antics relacionats amb l'àmbit. Per la segona vessant, vam considerar recopilar corpus de diverses fonts, com per exemple, de bases de dades de corpus que es troben a internet, com per exemple d'Opus Corpus[1] o de bases de dades de corpus biomèdics, com per exemple de The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations (2018).

Finalment, vam poder extreure tots els corpus externs de la segona font, on també s'hi trobaven corpus que apareixien a la primera. D'allà ens vam centrar, sobretot, en el corpus del MedlinePlus, que es un servei d'informació en línia que posa a disposició la NLM (National Library of Medicine) d'Estats Units, que té corpus en anglès i espanyol sobre temes de salut, de medicaments, proves de laboratori i enciclopèdia mèdica.

Tot i haver fet tota la cerca i descàrrega, al final del procés vam comprovar que començaríem l'entrenament del motor només amb els corpus de l'empresa i més endavant, si calia, hi afegiríem la resta de corpus o part d'ells. Per això, a partir d'ara totes

[1] Opus Corpus: http://opus.nlpl.eu/

les explicacions aniran enfocades als corpus interns, tot i que el procés seria el mateix pels externs.

Un cop havíem descarregat el corpus, havíem de procedir a preparar-los de cara al procés d'entrenament del motor.

4.4.1. Canvi de format de TMX a text pla tabulat

Donat el fet que tots els corpus interns de l'empresa ens van arribar en format TMX (Translation Memory eXchange), vam fer un primer pas on vam convertir les memòries a format de text pla tabulat, és a dir a format .txt tabulat. En aquell moment, vam ser conscients que hi havia diverses manera de convertir-les, però vam voler experimentar amb una eina nova: el TMXEditor, un programa gratuït que simplement va requerir una subscripció via correu electrònic, per tal d'obtenir un codi d'accés. Aquesta eina ens oferia moltes accions a fer, però simplement vam optar per canviar les llengües de l'arxiu original, ja que com que es tractava de diversos arxius, les llengües portaven una nomenclatura diferent, la qual cosa ens podria causar problemes, així que vam decidir canviar totes les llengües a EN com a llengua original i ES com a llengua meta. També, evidentment, vam convertir aquests arxius de TMX a TXT tabulat. Observem aquest procés a l'Annex 3.

Tot i així, també vam experimentar i fer el mateix procés de passar de TMX a text pla tabulat amb l'script que se'ns oferia als programes de MTUOC. Per a fer-ho simplement vam haver d'introduir les següents indicacions a la línia de comandes i l'arxiu es convertia directament. Les indicacions són:

```
python3 MTUOC_TMX2tabtxt.py -i corpus1.tmx -o corpus1.txt -s en -t es
```

Vam observar, però, que algunes de les memòries s'havien creat a partir de SDL Trados Studio i, per tant, els codis de llengua no eren els correctes. Simplement a l'hora de passar les indicacions mostrades anteriorment, havíem de seguir -s i -t amb la nomenclatura de la memòria, per exemple:

```
python3 MTUOC_TMX2tabtxt.py -i corpus1.tmx -o corpus1.txt -s en-US -t es-x-mo-SDL
```

4.4.2. Preparació dels corpus en un únic arxiu

És imprescindible que tots els corpus quedin, abans de l'entrenament, en un sol arxiu, així que hem de passar uns processos per tal d'unir-los. Per a fer-ho, simplement hem de tenir tots els corpus que haguem descarregat o obtingut en una mateixa carpeta i amb la mateixa combinació de llengües i passar la següent indicació al terminal:

```
cat A B C | sort | uniq | shuf > corpustots.txt
```

En aquesta indicació, les lletres A, B i C, corresponen al nom d'arxiu dels corpus que tinguem i "corpustots.txt" correspon al nom d'arxiu que li posem a l'arxiu de sortida, on hi tindrem tots els arxius. Aquesta acció, a més, no només uneix tots els arxius en un de sol, sinó que també elimina les repeticions que hi ha entre aquest arxius i ordena les oracions del corpus aleatòriament.

4.4.3. Separació de corpus per llengües

Si bé és veritat que volem entrenar el nostre motor amb corpus paral·lels, és possible que per a algun dels processos de preparació dels corpus necessitem tenir el corpus en dos arxius separats per llengües, és a dir, un arxiu amb totes les oracions en anglès i un arxiu amb totes les oracions en espanyol.

Per a fer-ho, és tan senzill com passar una altra indicació al terminal i en un moment ho tindrem fet. Recordem que a l'apartat anterior hem ajuntat tots els corpus en un arxiu que hem anomenat "corpustots.txt". Ara treballarem sobre aquest arxiu. La indicació és la següent:

```
cut -f 1 corpustots.txt > corpus_en
```

```
cut -f 2 corpustots.txt > corpus_es
```

D'aquesta manera, ens trobem amb dos corpus resultats, el "corpus_en" que és el corpus de les oracions en anglès i el "corpus_es" que és l'arxiu amb les oracions en espanyol.

4.4.4. Neteja de corpus

Tan bon punt vam tenir els corpus en el format que necessitàvem, vam procedir a un primer procés de neteja de corpus. Com ja sabem, és imprescindible que els corpus estiguin "nets", és a dir, que no hi hagi elements que puguin alterar l'entrenament i el funcionament del nostre motor. Per a fer tot això, MTUOC ens proporciona un programa amb diversos scripts que permeten dur a terme diversos accions (MTUOC_clean_parallel_corpus.py). Els elements esmentats i les accions que es poden dur a terme són les següents: l'apòstrof tipogràfic, que hem de reemplaçar per l'estàndard; les etiquetes html i xml, que o bé s'han d'eliminar o bé s'han de substituir pels caràcters que representen; eliminar els segments paral·lels on alguna de les dues llengües d'entrenament tingui un segment buit, els segments on el text de la llengua de partida i la d'arribada sigui igual i els segments on alguna de les dues llengües té menys de 5 caràcters o té un percentatge de números superior al 60%, i verificar que els codis de les llengües siguin els adequats, seguint els codis ISO de dues lletres. També hi ha uns arxius que indiquen uns elements (stringFromFile) o unes expressions regulars (regexFromFile) i si hi ha segments que els contenen, s'eliminaran.

Totes aquestes opcions, excepte la verificació dels codis de llengua i les dues últimes accions explicades, es poden dur a terme a la vegada, en un mateix procés i això és el que vam fer nosaltres. Per a fer-ho, vam escriure les indicacions següents a la línia de comandes de l'ordinador:

```
python3 MTUOC_clean_parallel_corpus.py -i corpustots.txt -o corpustots_net.txt -a
```

Amb això, vam poder eliminar tots aquells elements que malmetrien el nostre entrenament, sobretot les etiquetes, ja que molts dels corpus n'estaven plens. A la documentació d' Oliver, A. (2020a) hi apareixen també les comandes per executar els processos de neteja per separat.

4.5. Preprocessament de corpus

Abans de començar la fase de l'entrenament i, doncs, una de les més importants, vam haver de confirmar que disposàvem de totes les eines necessàries, pel que fa a software i hardware. Un cop aquest aspectes estaven preparats, vam poder procedir a l'entrenament.

Després del recull de tots els corpus, vam obtenir un total de poc més d'un milió de segments alineats de corpus interns de l'empresa, concretament 1.132.912.

A continuació, explicarem tots els processos que s'inclouen a la preparació del corpus que hem agrupat a l'apartat 4.4.2. Preparació dels corpus en un únic arxiu. Cal dir que tots aquests processos es poden dur a terme amb una sola acció gràcies a la preparació que ens ensenya Oliver (2020a). Aquesta única acció simplement es duu a terme amb el corpus que hem agrupat i que hem anomenat "corpustots.txt" i amb els arxius que es descarreguen de MTUOC. A partir d'aquí, passem la indicació següent i obtenim el resultat dels processos que explicarem a continuació:

```
bash MTUOC-preprocess.sh corpustots_net.txt
```

Amb aquesta acció, se'ns creen una sèrie d'arxius:

corpus.BPE.es	corpus.tok.es	train.en
corpus.BPE.en	corpus.true.en	train.es
corpus_en	corpus.true.es	val.BPE.en
corpus_es	eval.BPE.en	val.BPE.es
corpus.split.en	eval.BPE.es	val.en
corpus.split.es	eval.en	val.es
corpus.tok.clean.en	eval.es	vocab_BPE.en
corpus.tok.clean.es	train.BPE.en	vocab_BPE.es
corpus.tok.en	train.BPE.es	

No oblidem, però, que tots aquest arxius provenen de diferents processos que s'han dut a terme a partir d'aquesta acció. És per això, que considerem important mostrar tots els processos per tal d'entendre'ls i poder-los dur a terme per separat si en alguna circumstància ens és necessari. Tots aquests processos, doncs, s'expliquen a continuació.

Com hem esmentat, el procés de preprocessament dels corpus es basa en diverses accions, totes elles imprescindibles per tenir els nostres corpus preparats per entrenar el motor. Aquestes accions, o la gran majoria d'elles, es duen a terme tant per l'entrenament de motors de traducció automàtica estadística com per motors neuronals. Les principals accions que hem dut a terme són les següents: la tokenització, el truecasing, la neteja de corpus, el tractament de les expressions numèriques i el càlcul de subwords i la divisió del corpus en entrenament, validació i avaluació. El càlcul de subwords és una acció específica de l'entrenament de motors neuronals.

4.5.1. Tokenització

Com ja hem comentat a l'apartat 3.1.5. Traducció automàtica neuronal (TAN), el procés de tokenització es basa en la separació de les unitats lèxiques. En les nostres dues llengües de treball, l'espanyol i l'anglès, aquesta separació ja existeix de normal, però en aquest procés es separen també les paraules dels signes de puntuació que les acompanyen. És un procés de vital importància.

A l'exemple de l'Annex 4, mostrem el resultat d'aquest procés. Podem tant passar una oració com un arxiu sencer; en el nostre cas, hi passem l'arxiu que hem dividit anteriorment a l'apartat 4.4.3. Separació de corpus per llengües en "corpus_en" i "corpus_es". Fem servir el tokenitzador d'anglès per a un arxiu en anglès i un tokenitzador en espanyol pel mateix arxiu en espanyol.

Per al tokenitzador en anglès, necessitem la indicació següent en el nostre terminal, tenint en compte que l'arxiu d'origen s'anomena "corpus_en" i el de destí "corpus.tok.en":

```
python3 MTUOC_tokenizer_eng.py < corpus_en > corpus.tok.en
```

El mateix per a la tokenització en espanyol, tenint en compte que els noms dels arxius i del programa canvien:

```
python3 MTUOC_tokenizer_spa.py < corpus_es > corpus.tok.es
```

El resultat d'aquesta acció l'observem a l'Annex 4, on podem comprovar que hi ha hagut una separació entre les paraules i els signes de puntuació, en el cas de les nostres llengües.

4.5.2. Truecasing

Com hem comentat a l'apartat 3.1.5. Traducció automàtica neuronal (TAN), el procés del truecasing assigna a cada token de les majúscules i minúscules tal i com li corresponen.

Aquest procés requereix un pas més que l'anterior, ja que abans de dur-lo a terme s'ha d'entrenar un model de truecasing a partir del corpus que tinguem i d'un diccionari (si s'escau). El procés d'entrenament d'un model de truecasing es fa de manera separada per cada llengua, així que tornarem a fer ús dels corpus per llengües, tot i que aquesta vegada ho farem amb els corpus que ja hem tokenitzat. A continuació, mostrem la indicació del terminal que durà a terme el procés d'entrenament de models de truecasing. Aquesta indicació consta de 4 parts: l'arxiu .py que duu a terme el procés; el nom del model que volem crear, que segueix la indicació -m; el corpus que hem tokenitzat anteriorment, que segueix la indicació -c, i el diccionari (si en tenim) de la llengua que tractem, que segueix la indicació -d. Així doncs, així serien les indicacions per entrenar el model en anglès i el model en espanyol:

```
python3 MTUOC_train_truecaser.py -m tc.en -c corpus.tok.en -d eng.dict
```

```
python3 MTUOC_train_truecaser.py -m tc.es -c corpus.tok.es -d spa.dict
```

Tan bon punt s'han creat els nostres models, ja podem dur a terme el procés de truecasing dels nostres corpus tokenitzats. A la següent indicació hi veiem l'arxiu .py, el nom del model que hem creat al pas anterior, el corpus tokenitzat i finalment el nom que li assignem al corpus resultant de l'acció, tant en anglès com en espanyol.

```
python3 MTUOC_truecaser.py tc.en < corpus.tok.en > truec_en
```

```
python3 MTUOC_truecaser.py tc.es < corpus.tok.es > truec_es
```

Amb això ja tenim els nostres corpus tokenitzats i amb les majúscules i minúscules tal i com corresponen, com veiem a l'Annex 5, on podem comprovar que, per exemple, en anglès la paraula "Lornaxicam" s'ha mantingut amb la majúscula, mentre que en espanyol aquesta majúscula s'ha eliminat. Això es deu a que tant al diccionari que hem fet servir per entrenar el model en anglès com a tot el corpus, aquesta paraula surt escrita en majúscula i, per tant, el nostre model ha après que s'ha de mantenir, mentre que el model espanyol ha après que s'ha d'eliminar. Trobem la paraula destacada a l'Annex 5.

4.5.3. Neteja de segments llargs

La neteja de segments llargs es duu a terme per una raó: hi ha vegades en què segments molt llargs poden confondre el procés d'entrenament del nostre motor, així que és preferible eliminar-los. A l'hora de fer-ho, som nosaltres mateixos el qui decidim quina mida màxima volem que tinguin els nostres segments. En el nostre cas, decidim que volem que la llargada màxima dels nostres segments sigui de 80 tokens. Hem de tenir en compte que, tal i com acabem de veure, el número que posem és el número de tokens que volem i, per tant, l'arxiu que processarem és l'arxiu amb els corpus tokenitzats.

Per tant, de la mateixa manera que en els apartats anteriors, a continuació mostrem les indicacions que hem d'inserir a la línia de comandes per poder executar aquesta acció:

```
python3 MTUOC_cleaning.py corpus.tok en es 80
```

Amb aquesta indicació se'ns netegen tant el corpus "corpus.tok.en" com el corpus "corpus.tok.es" i se'ns creen els arxius "corpus.tok.clean.en" i "corpus.tok.clean.es".

Observem novament aquest procés a l'Annex 6.

4.5.4. Tractament d'expressions numèriques

Un altre procés important és el tractament de les expressions numèriques dels nostres corpus. Necessitem adaptar els nombres a la forma que el nostre motor els entengui i els pugui processar. Existeixen dues maneres de fer-ho que es diferencien pel fet que una està destinada a traductors automàtics estadístics i l'altra a traductors automàtics neuronals. Com que nosaltres estem creant un TA neuronal, parlarem només d'un dels dos programes. L'altre es troba explicat amb tota la documentació d'aquests processos gràcies a Oliver, A. (2020a).

En el cas del programa de tractament d'expressions numèriques de TA neuronals que ens ofereix MTUOC, separa les xifres de les expressions numèriques amb espais en blanc en el nostre corpus d'entrenament. També ho farà en el moment en què introduïm una oració per a traduir al nostre TA i després tornarà a col·locar els números tal i com estaven. De la primera tasca se n'encarrega el programa MTUOC_splitnumbers.py i de la segona, el programa MTUOC_desplitnumbers.py. En el nostre cas, però, com que ens interessa

separar-los per a preparar els nostres corpus d'entrenament, només parlarem del primer programa.

Igual que hem fet amb la resta de processos, introduïm les indicacions al terminal per tal d'executar el programa, com veurem a continuació. L'arxiu que decidim processar és el "corpus.true" tant en anglès com en espanyol i volem que el resultat s'anomeni "corpus.split.en/es" respectivament en cada idioma.

```
python3 MTUOC_splitnumbers.py < corpus.true.en > corpus.split.en
```

```
python3 MTUOC_splitnumbers.py < corpus.true.es > corpus.split.es
```

Volem recordar que, tot i que en el nostre cas hem decidit que el corpus a processar per aquest programa seria el "corpus.true", és a dir, el corpus que ha passat pel procés de truecasing, podem fer-ho sinó també amb el que ha estat tokenitzat o amb un altre arxiu de corpus que es trobi en el format de text pla tabulat. Tot el procés es pot observar a l'Annex 7.

4.5.5. Divisió dels corpus en entrenament, validació i avaluació

Aquest procés és imprescindible a l'hora d'entrenar un traductor automàtic neuronal. Es tracta de dividir el corpus que tenim en tres arxius diferents: un arxiu on hi ha la major part dels segments del corpus per a entrenar, un arxiu d'uns 1.000 segments per fer la validació després de cada entrenament i un arxiu de referència d'uns 1.000 segments també, per a què el nostre traductor faci una avaluació automàtica cada vegada que s'acabi d'entrenar o quan es desitgi (que es pot veure a l'apartat 4.7. Avaluació).

Aquest procés el podem dur a terme gràcies al programa MTUOC_split_corpus.py. Abans d'executar-lo, però, haurem de saber el nombre de segments que té el nostre arxiu de corpus que volem dividir. Per a fer-ho, simplement podem obrir el nostre corpus amb un editor de textos qualsevol i anar a la última línia per comprovar quantes n'hi ha. En el cas dels arxius que fem servir com a exemple, hi ha tant a l'anglès com a l'espanyol, 695.175 segments, així que els haurem de dividir en:

- 693.175 segments d'entrenament, als que anomenarem train.en/es respectivament per cada idioma;

- 1.000 segments de validació, als que anomenarem val.en/es, respectivament per cada idioma;

- 1.000 segments d'avaluació, als que anomenarem eval.en/es, respectivament per cada idioma.

Així doncs, executem el programa de la següent manera i n'observem els resultats:

```
python3 MTUOC_split_corpus.py corpus_en train.en 693175 val.en 1000 eval.en 1000
```

I fem el mateix pel corpus en espanyol:

```
python3 MTUOC_split_corpus.py corpus_es train.es 693175 val.es 1000 eval.es 1000
```

Com era d'esperar, d'aquestes dues accions obtenim 6 arxius, que seran imprescindibles pels processos d'entrenament, validació i avaluació.

4.5.6. Càlcul de subwords

A l'hora de dur a terme aquesta fase, seguim fixant-nos en les instruccions que ens proporciona Oliver, A. (2020a). Per tant, si el que volem fer és aplicar el *byte-pair-encoding* subwords als nostres arxius indicarem a línia de comandes la instrucció que mostrem a continuació. Hem de tenir en compte que aplicarem aquest càlcul als arxius: corpus_en, corpus_es, train.en, train.es, val.en, val.es, eval.en i eval.es.

```
python3 MTUOC_BPE.py apply codes_file < corpus_en > corpus.BPE.en
```

Com hem dit, introduïm aquesta instrucció amb els arxius esmentats per tal d'obtenir els corpus llestos per l'entrenament.

A l'Annex 8, podem observar, com amb la resta d'apartats, el terminal amb les indicacions esmentades i, a continuació, mostrem el contrast entre l'arxiu "corpus_en" i l'arxiu "corpus.BPE.en", on podrem veure clarament com es marca el *byte-pair-encoding*.

4.6. Entrenament

A l'hora de posar-nos a entrenar el motor, vam haver de tenir en compte el principal factor que era el hardware de què disposàvem. És evident que per poder entrenar un motor de qualitat necessites potència, per tal també de no trigar molt de temps en dur a terme aquest entrenament. Per tant, és imprescindible disposar d'una o més GPU (Graphical Processing Unit), per tal de dur a terme un entrenament exitós i mitjanament ràpid, ja que sense GPU l'entrenament pot arribar a ser etern o durar anys i, per tant, és impossible.

Si bé es veritat que pel volum de corpus que disposàvem necessitem fer un entrenament gran, existeix la possibilitat de fer un entrenament més senzill com a prova, que ens permet també veure el procediment que es segueix. Per tant, a continuació explicarem com seria l'entrenament d'un sistema bàsic i hi afegirem alguna explicació sobre entrenar sistemes més complexes.

Primerament, exposem com dur a terme un entrenament senzill com a prova, tot extret de la informació que enes proporciona Oliver (2020b). Recordem, que després de tots els processos anteriors hem obtingut una sèrie d'arxius. Per posar en marxa un entrenament senzill necessitem els arxius: train.BPE.en, train.BPE.es, val.BPE.en i val.BPE.es. A partir d'aquí, la indicació que posem a la línia de comandes és la següent:

./marian --train-sets train.BPE.en train.BPE.es –valid-sets val.BPE.en val.BPE.es --vocabs vocab-en.yml vocab-es.yml --model model.npz

A partir d'això, com veiem, se'ns generen els arxius vocab-en.yml, vocab-es.yml i model.npz. Aquests arxius composen el nostre traductor automàtic.

Quant al nostre entrenament, la mecànica ha sigut diferent a l'explicada. Per dur a terme l'entrenament, hem fet servir un únic arxiu que ja incorporava tota la informació que hem aportat a l'exemple anterior, amb els mateixos arxius i altres informacions addicionals. L'arxiu s'ha anomenat config-s2s-bleu.yml i la indicació que s'ha de posar al terminal és la següent:

marian -c config-s2s-bleu.yml

A partir d'aquí, es fa també una validació a partir de l'arxiu MTUOC-bleu.py i l'arxiu MTUOC-validate.sh. Aquesta validació, com veiem, es fa a partir de BLEU, però no és

un BLEU real, perquè es fa abans de desfer el BPE de les paraules i, per tant, no són les paraules reals en sí, sinó les paraules separades en subwords.

L'entrenament del nostre motor va durar, aproximadament, un dia i mig. En aquest transcurs de temps, el motor validava les traduccions que anava fent i mostrava el BLEU de la que ell considerava millor i així successivament. L'entrenament va acabar després d'arribar al punt que nosaltres havíem marcat com a *early_stopping*, és a dir, a la configuració de l'entrenament, vam indicar que si en 5 processos de validació no es millorava el BLEU, l'entrenament s'atura i s'agafa el model que ha obtingut el BLEU més alt fins llavors. Com hem dit abans, no és un BLEU real i a continuació veurem que l'últim BLEU va ser de 0,77, mentre que el BLEU real que hem obtingut amb l'avaluació és completament diferent, però això es pot comprovar a l'apartat següent. A continuació, veiem un diagrama on observem l'evolució del BLEU durant el procés d'entrenament i validació del motor. Aquest diagrama es basa en un arxiu que es crea durant aquesta validació, anomenat valid.log, les dades del qual podem observar a l'Annex 9.

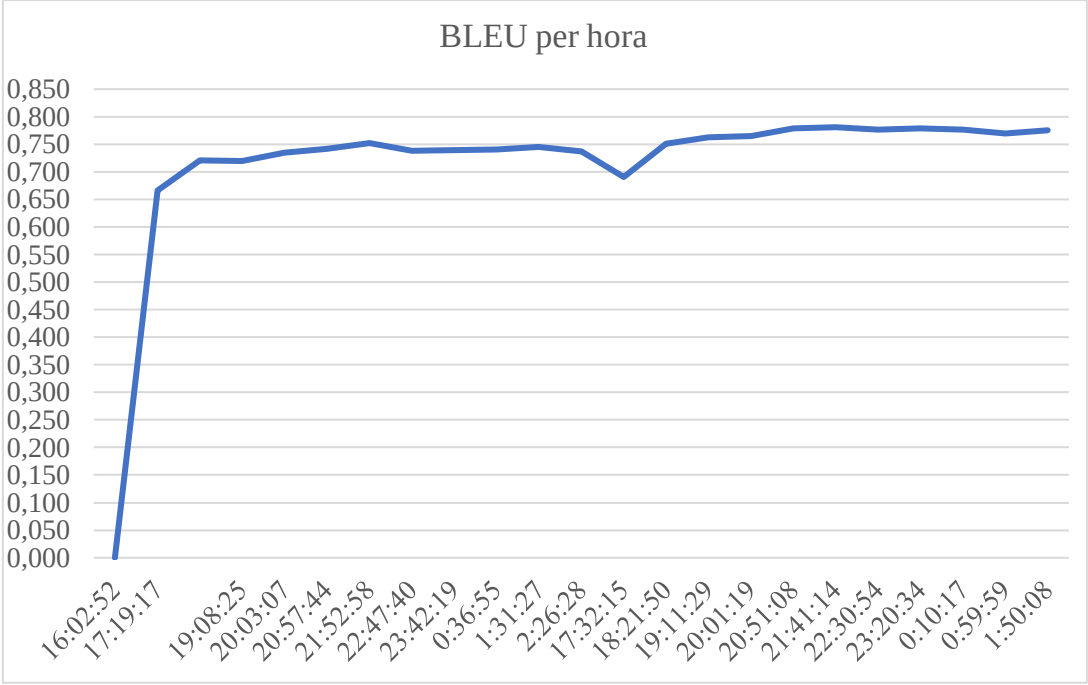


Figura 4. Diagrama de BLEU durant la validació

Observem que durant el procés de validació el BLEU augmenta o disminueix a mesura que avança i, a mesura que analitza més dades, el BLEU es manté en una puntuació

bastant similar. A les dades que comprovem a l'Annex 9, podem veure el procés de manera més clara i també la seva durada.

A partir d'aquí, ja tindrem el nostre motor entrenat i podrem passar a una de les últimes fases, que és l'avaluació, amb la qual podrem observar els resultats del nostre traductor i si aquests són positius o si caldrà dur a terme un altre entrenament amb dades diferents o amb més dades per millorar-lo.

4.7. Avaluació

A l'hora de dur a terme l'avaluació, vam decidir que era oportú fer-ne de diferents tipus. Els dos principals mecanismes d'avaluació de traducció automàtica són l'avaluació quantitativa i l'avaluació qualitativa. Pel que fa a la quantitativa, vam decidir avaluar automàticament a partir de les mètriques esmentades a l'apartat 3.2. Avaluació i mètriques, mentre que per l'avaluació qualitativa ens vam decantar per la categorització d'errors que els posteditors van trobar en els resultats del nostre traductor automàtic i un rànquing de comparació amb altres motors de TA. Aquest dos processos s'expliquen als apartats següents.

4.7.1 Avaluació quantitativa

Per dur a terme l'avaluació del nostre sistema de traducció automàtica, farem referència a les mètriques esmentades a l'apartat 3.2. Avaluació i mètriques, les mètriques BLEU, NIST, WER i TER. Per a fer-ho, obtindrem com a text de referència els 1.000 segments que, a l'hora del preprocessament dels corpus, s'han apartat per dur a terme l'avaluació (apartat 4.5.5. Divisió dels corpus en entrenament, validació i avaluació). Aquests 1.000 segments alineats formaven part dels corpus que hem recollit per preparar l'entrenament del nostre motor, però han estat exclosos de l'entrenament per poder-se usar en aquest pas.

A l'hora d'avaluar la qualitat de manera automàtica, hem fet servir també un dels scripts que ens proporciona Oliver, A. (2020a) i l'hem introduït a la nostra línia de comandes, de la mateixa manera que hem fet amb els altres scripts. En aquest cas, comptem també amb

interfície gràfica per dur a terme l'avaluació automàtica. Tot això ho mostrem a l'Annex 10. Per tal de no només obtenir els nostres resultats, hem decidit també fer una mateixa avaluació de l'arxiu eval.en amb un dels traductors per excel·lència del mercat, Google Translate. Per a fer-ho, hem seguit els mateixos passos que amb el nostre traductor: hem traduït el document, si no hem obtingut un arxiu de text pla ho hem passat a aquest format, i l'hem passat per l'script d'avaluació.

Com hem dit anteriorment, ho fem a partir d'un script d'avaluació. Prèviament, però, ens hem hagut de descarregar la carpeta amb els scripts que ens facilita Oliver, A. (2020a) i a dins d'aquella carpeta, amb els nostres arxius també, farem:

python3 MTUOC-eval.py --tokenizer MTUOC_tokenizer_spa.py --refs eval.es --hyp eval_motor.es.hyp

I el mateix per l'arxiu amb les traduccions de Google Translate:

python3 MTUOC-eval.py --tokenizer MTUOC_tokenizer_spa.py --refs eval.es --hyp eval.Google.es.hyp

Compararem, doncs, els resultats que ens proporciona el nostre motor, juntament amb el traductor automàtic genèric de Google. Aquestes dades s'analitzaran a l'apartat 5. Resultats.

	Motor creat	**Google Translate**
BLEU	0,4061	0,3445
NIST	8,217	7,459
WER	0,519	0,6402
EDIT DISTANCE	37,08%	44,0%
TER	0,436	0,56

Taula 3. Resultats de l'avaluació automàtica quantitativa

4.7.2. Avaluació qualitativa

A l'hora de dur a terme l'avaluació qualitativa i, per tant, l'avaluació manual, hem hagut de tenir en compte alguns factors. Primerament, hem hagut de comprovar que disposàvem de tots els recursos necessaris, sobretot els humans. A partir de l'empresa amb qui hem treballat hem pogut disposar de 8 traductors i revisors que han participat en una de les dues tasques d'avaluació. Per l'altra tasca, s'ha seleccionat un grup de 12 traductors i posteditors. A continuació presentem les tasques mencionades.

4.7.2.1. Categorització d'errors

La primera tasca que s'ha dut a terme per avaluar la qualitat del nostre traductor automàtic ha sigut l'anàlisi i categorització d'errors a partir d'una plantilla d'errors. Per fer-ho, hem presentat 4 fragments de textos de 39 segments cada un i els hem traduït amb el nostre motor. A partir d'aquí, hem incorporat els textos a l'eina de traducció assistida per ordinador memoQ, dues vegades cada text per poder-lo assignar a dos correctors diferents. A més a més, memoQ compta amb diversos models de categorització d'errors que pots incorporar en el teu projecte per tal de dur a terme una tasca com la que volem nosaltres. Té els models més establerts en l'actualitat com són el DQF de TAUS o el model de LISA i un model que proposa la pròpia eina. Nosaltres, però, hem decidit fer ús d'un model adaptat. Juntament amb l'empresa amb qui hem fet un projecte, hem decidit fer ús del model actualitzat que fan servir actualment els seus revisors per dur a terme les revisions, per fer més fàcil la tasca als posteditors, ja que tots hi estan familiaritzats.

El model que hem fet servir, consta de 9 categories amb les seves respectives subcategories. Les categories són: *meaning, language, terminology, style, layout, country standards, repeat, kudos* i *other.* A part de seleccionar la categoria i subcategoria a què pertany cada error, els posteditors també han atribuït una gravetat a aquest error. Les gravetats es dividien en 4 nivells, que són: *critical, major, minor* i *preferential/improvement*. A continuació, mostrem una taula amb totes les categories i subcategories explicades, seguida d'una taula amb les quatre gravetats degudament explicades també.

Error type	Subcategory	Explanation
Meaning		El text d'arribada no reflecteix correctament el text de partida. Hi ha diferències de significat i mala traducció.
	Addition	La traducció inclou text que no apareix a l'original.
	Omission	Hi ha contingut que apareix a l'original que no està present a la traducció.
	Mistranslation	La traducció no representa correctament el contingut de l'original.
	Ambiguous	La traducció és ambigua, mentre que l'original no ho pretén ser.
	Not appropriate for context	El significat que transmet la traducció no és apropiat en aquest context.
	Untranslated text	Hi ha contingut que s'havia de traduir i s'ha mantingut en la llengua original.
	Word/Tag order	Les paraules o les etiquetes estan en un ordre que no toca.
	Other	Altres errors de significat.
Language		Errors relacionats amb la forma o el contingut del text, deixant de banda si és una traducció o no, és a dir, si el text és correcte lingüísticament.
	Punctuation	La puntuació s'ha fet servir de manera incorrecta.
	Spelling	Errors relacionats amb l'ortografia.
	Grammar-Syntax	Errors relacionats amb la gramàtica o la sintaxi del text.
	Capitalization	Errors relacionats amb l'ús de majúscules i minúscules.

	Other	Altres errors relacionats amb la forma lingüística del text.
Terminology		Un terme (una paraula d'un àmbit específic) s'ha traduït per un altre terme que no és el correcte pel domini o àmbit en què es presenta.
	Inconsistent	Terminologia inconsistent en el text. Aquest terme s'ha traduït d'una altra manera anteriorment.
	Glossary non-compliance	El terme no es correspon amb l'indicat al glossari.
	TM non-compliance	El terme no es correspon amb l'indicat a la memòria de traducció.
	Prev.correction non-compliance	El terme no es correspon amb la correcció prèviament feta.
	Industry non-compliance	El terme no es correspon a l'àmbit de la indústria del text.
	3rd party non-compliance	El terme no és propi de la indústria d'un tercer.
	Not appropriate for context	La terminologia no és l'apropiada pel context del text.
	Other	Altres errors relacionats amb els termes específics del text.
Style		El text té errors estilístics.
	Awkward syntax	El text està escrit en un estil complicat.
	Non-compliance with SG	L'estil no es correspon al marcat a la guia d'estil.
	Inconsistent with refmat	L'estil del text no és consistent amb el material de referència.
	Inconsistent within text	L'estil del text no és consistent en si.

	Unidiomatic use	El contingut és gramatical però no és idiomàtic.
	Tone	No respecta el to del text. El to no és l'adequat.
	Literal translation	És una traducció literal del text original i, per tant, no correspondria al text d'arribada.
Layout		Hi ha errors en la disposició i el disseny del text.
	Formatting	El format del text no és correcte.
	Corrupted tags	Alguna etiqueta s'ha modificat i ja no és vàlida.
	Missing tags/variables	Falten etiquetes o variables.
	Link not working	L'enllaç no funciona.
	Truncation/Overlap	El text queda truncat (queda tallat i no es veu completament) o superposat.
	String-length error	Error en l'allargada del segment. El segment és molt més llarg o curt que l'original.
	Missing/Invisible text	Falta text o s'ha configurat com a invisible.
	Corrupted characters	Algun caràcter s'ha modificat i no és vàlid.
	Incorrect cross-ref	Referència creuada incorrecte.
Country Standards		El text no s'adhereix a les convencions mecàniques del local i no respecta els requisits de presentació de contingut propis de la llengua meta.
	Dates, units of measurement, addresses, phone numbers, zip codes, shortcut keys,	Les dates, unitats de mesura, adreces, números de telèfon, etc. estan escrits en un format que no és el propi de la llengua i cultura d'arribada. També inclou qualsevol

	cultural references…	referència cultural que no s'ha aplicat a la llengua d'arribada.
Repeat		L'error es repeteix.
	Repeated error	S'assigna a tots els errors que han aparegut repetits i que es considera que no han de tornar a penalitzar.
Kudos		Premis
	😊 **Congratulations**	S'atribueix als segments que són realment molt bons i sorprenen al revisor/posteditor.
Other		Qualsevol altre tipus d'error no especificat.

Taula 4. Categorització d'errors

Critical	Errors que poden comportar implicacions de salut, de seguretat, legals o financeres, transgredir guies d'ús geopolítiques, perjudicar la reputació de l'empresa, fer que l'aplicació no funcioni o modificar/representar negativament la funcionalitat d'un producte o server, o que pot semblar ofensiu.
Major	Errors que poden confondre o enganyar a l'usuari o dificultar l'ús correcte d'un producte/servei degut a un canvi significant en el significat o per errors que apareixen a una part visible o important del contingut.
Minor	Errors que no porten a la pèrdua de significat i no confonen a l'usuari o dificulten l'ús però es poden veure, empitjoren la qualitat estilística, la fluïdesa o claredat, o fan que el contingut sigui menys atractiu.
Preferential/ Improvement	S'usa per categoritzar informació original, problemes o canvis que es poden fer però no compten com a errors. Moltes vegades representen una preferència del revisor o una millora en l'estil, errors repetits, o canvis del glossari o les instruccions que encara no s'han implementat, és a dir, un canvi que s'ha de fer però del qual no s'ha informat al traductor.

Taula 5. Gravetats dels errors

4.7.2.2. Rànquing de traductors automàtics

La segona tasca d'avaluació que hem dut a terme és la classificació de traductors automàtics. Per aquesta tasca, hem comptat amb l'ajuda de 8 traductors i posteditors. Hem dut a terme la tasca a partir de la plataforma de DQF de TAUS[2], tot creant-nos un usuari, que ens ha permès preparar la tasca amb una gran facilitat i poder-la desenvolupar d'una manera molt senzilla.

La tasca ha consistit en la recopilació de 100 segments d'un text, la traducció d'aquests 100 segments amb tres motors de traducció automàtica diferent, la preparació de l'arxiu i l'avaluació dels traductors. Explicarem a continuació, tot el procés de preparació i avaluació pas per pas.

Primerament, com ja hem dit, hem hagut de triar 100 segments d'un text de referència. En aquest cas, vam utilitzar textos d'un parell de corpus biomèdics que hi ha a internet. A partir d'aquí, hem agrupat tots els segments en anglès a un arxiu de text pla, i els hem traduït amb tres motors diferents. El primer motor és Google Translate, el segon DeepL i el tercer, evidentment, és el que hem creat nosaltres al qual hem anomenat ara per ara *Marian*. Totes les traduccions les hem col·locat a un arxiu d'Excel amb les columnes que la plataforma DQF demanava. A continuació veiem una imatge de l'Excel, que podrem veure amb més detall també a l'Annex 11, on veurem els dos primers segments de l'arxiu i uns de més endavant ampliats:

ID	Source Segment	Segment Origin	Google	DeepL	Marian
1	Introduction: Multiple interventions are performed in critical patients admitted to Intensive Care Units (ICUs).	SampleText_1	Introducción: se realizan múltiples intervenciones en pacientes críticos ingresados en Unidades de Cuidados Intensivos (UCI).	Introducción: Se realizan múltiples intervenciones en pacientes críticos ingresados en las Unidades de Cuidados Intensivos (UCI).	Introducción: se realizan múltiples intervenciones en pacientes críticos ingresados en unidades de cuidados intensivos (UCI).
2	This study explores the presence in the daily practice of ICUs of elements related to the 6 bioethics quality indicators of the Spanish Society of Intensive and Critical Care Medicine and Coronary Units, and the participation of their members in the hospital ethics committees.	SampleText_1	Este estudio explora la presencia en la práctica diaria de las UCI de elementos relacionados con los 6 indicadores de calidad de bioética de la Sociedad Española de Medicina Intensiva y Crítica y Unidades Coronarias, y la participación de sus miembros en los comités de ética del hospital.	Este estudio explora la presencia en la práctica diaria de las UCI de elementos relacionados con los 6 indicadores de calidad bioética de la Sociedad Española de Medicina Intensiva y Crítica y de las Unidades Coronarias, y la participación de sus miembros en los comités de ética hospitalaria.	Este estudio explora la presencia en la práctica diaria de las UCI de elementos relacionados con los 6 indicadores de calidad bioética de la sociedad española de medicina intensiva y crítica y unidades coronarias, y la participación de sus miembros en los comités de ética hospitalaria.
	Materials and methods: A multicenter observational study was carried out, using a survey exploring descriptive aspects of the ICUs, with 25 questions related to bioethics quality		Materiales y métodos: se realizó un estudio observacional multicéntrico, utilizando una encuesta que explora aspectos descriptivos de las UCI, con 25 preguntas relacionadas con los indicadores de calidad de	Materiales y métodos: Se realizó un estudio observacional multicéntrico, mediante una encuesta que exploró los aspectos descriptivos de las UCI, con 25 preguntas relacionadas con los indicadores de calidad de	Materiales y métodos: se realizó un estudio observacional multicéntrico, utilizando una encuesta que explora aspectos descriptivos de las UCI, con 25 preguntas relacionadas con indicadores de calidad de la

Figura 5. Arxiu de preparació del rànquing

Un cop hem tingut aquest Excel llest, hem anat a la plataforma de DQF i hem preparat el projecte. Aquí, hem introduït un nom al projecte, hem seleccionat el tipus de projecte que volem, ja que la plataforma permet fer avaluacions de diversos tipus. En el nostre cas,

[2] https://dqf.taus.net/

hem triat “Comparison” i dins de “Comparison Type” hem triat “Rank comparison”, que permet als posteditors classificar de millor a pitjor les traduccions. Hem triat també el tipus de contingut que conté el text i la industria a què pertany el nostre projecte. Finalment, hem seleccionat la llengua de partida, l’anglès i la d’arribada, l’espanyol.

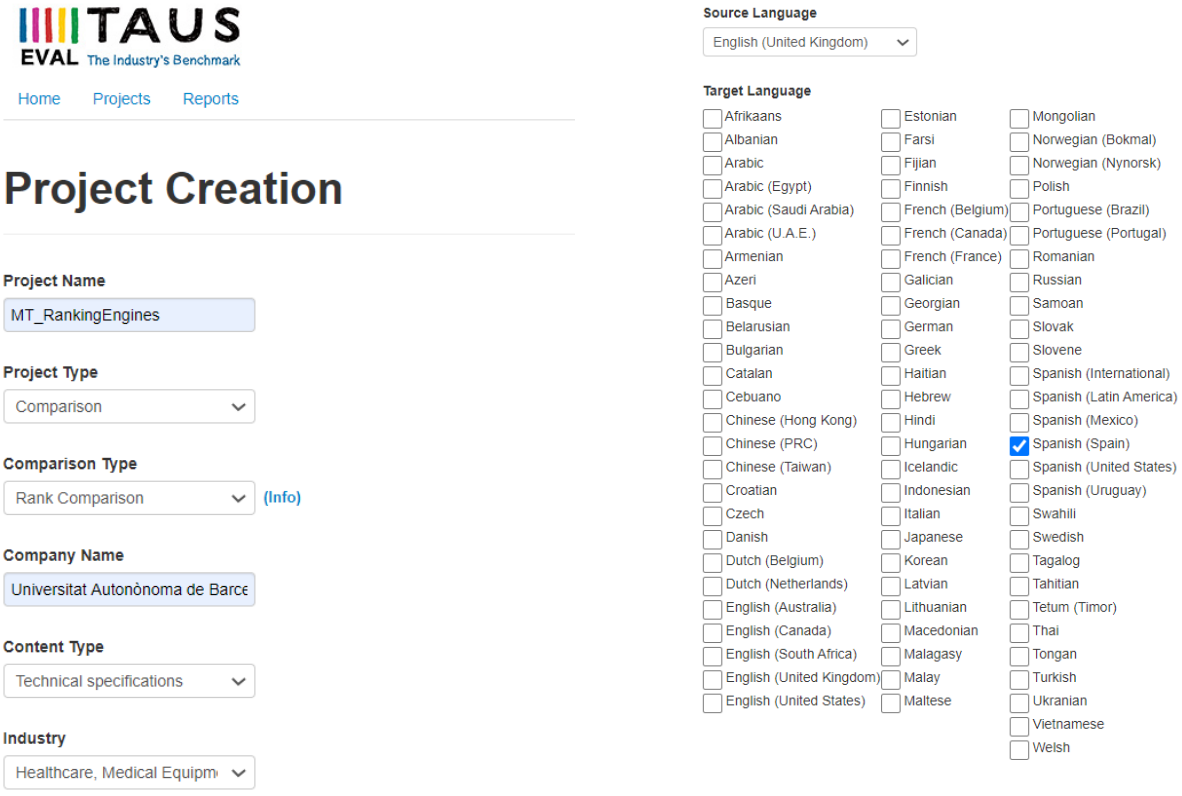


Figura 6. Creació del projecte de comparació

A continuació, la plataforma et demana que pugis l’arxiu en format .xls/xlsx o .csv i tornis a seleccionar les llengües. El programa, de fet, recomana que la comparació es faci amb 250 segments, però per qüestions de temps i logística, nosaltres hem optat per fer-la amb 100 segments, que és el mínim que ens permet el programa.

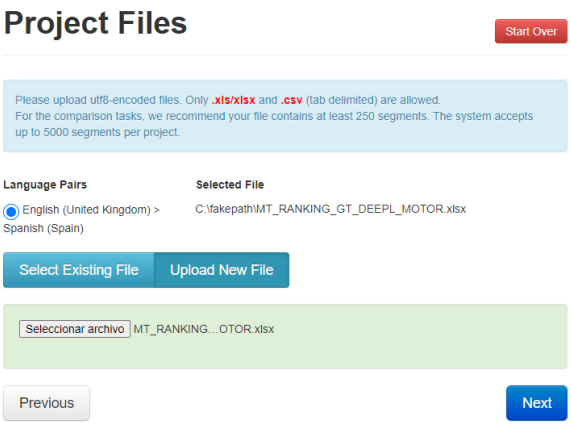


Figura 7. Selecció d'arxius per la comparació

A la següent pantalla, la plataforma et permet introduir el nom i el correu electrònic de tots els posteditors que participin a l'avaluació, tot i que més endavant ens permet també eliminar els posteditors que vulguem i afegir-ne també de nous. Per qüestions de privacitat de dades, no mostrarem el llistat amb els noms i correus electrònics dels posteditors, simplement la pantalla on podem introduir-los.

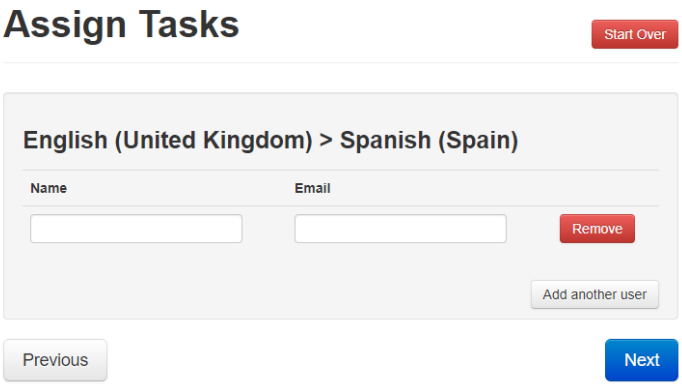


Figura 8. Assignació de tasques a traductors per la comparació

Finalment, et mostra una pantalla amb el resum de tota la informació que has anat introduint al llarg de la configuració i paral·lelament, el programa ha enviat per correu electrònic un enllaç als posteditors per poder accedir a l'avaluació. Un cop els avaluadors hi entren, l'aspecte de la pàgina web és el següent:

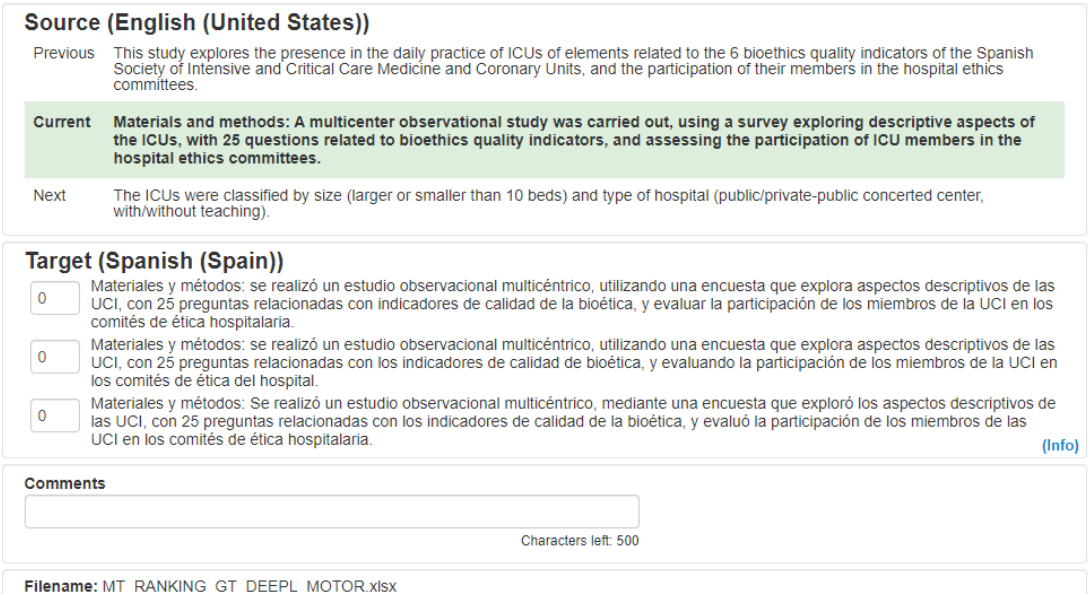


Figura 9. Pantalla de classificació de motors

El procés d'avaluació es basa, simplement, en assignar una puntuació de l'1 al 3 a les traduccions que se'ns ofereixen de cada segment, seguint la taula següent:

1	És la millor traducció de les tres que apareixen.
2	És la segona millor traducció de les tres que s'ofereixen.
3	És la pitjor traducció de les tres que apareixen.

Taula 6. Puntuacions per la classificació de motors

Un cop cada posteditor acaba l'avaluació, rebem un correu anunciant-ho i podem accedir a uns informes que veurem a l'apartat següent. També, a l'Annex 12 podem veure les instruccions que s'han enviat als posteditors per tal que puguin seguir amb facilitat el procés d'avaluació.

5. Resultats

En aquest apartat, exposem breument els resultats que hem obtingut dels tres processos d'avaluació que hem explicat a l'apartat anterior.

Quant a **l'avaluació automàtica**, només ens caldria dir que els números parlen per ells mateixos. És veritat que veiem unes xifres millors al nostre traductor que al de Google, la qual cosa crea unes grans expectatives quant a l'ús del motor. També és cert, però, que l'entrenament del nostre motor s'ha fet amb textos específics d'aquesta especialitat, mentre que el de Google està fet amb arxius genèrics. És normal que els resultats siguin millors quan es tracta de textos molt tancats d'un àmbit d'especialitat. Caldrà, doncs, veure els resultats de l'avaluació manual per comprovar si els resultats són tan positius com els que veuríem a l'automàtica. Hem de tenir en compte, sempre, que l'avaluació automàtica és una guia per veure el rendiment d'un motor, però mai ens podem conformar amb aquests resultats i cercar traductors per fer una avaluació humana.

Pel que fa a la **classificació dels errors**, un cop els posteditors han acabat la seva tasca de revisió, postedició i classificació dels errors, hem obtingut un informe amb tots els errors presentats en els 8 textos. En general, s'ha trobat un gran nombre d'errors als textos, la qual cosa era d'esperar també. Com podrem comprovar a la taula que mostrem a continuació, la majoria d'errors es concentren en el significat (*meaning*) i en l'estil (*style*), però veiem que la terminologia no presenta un gran nombre d'errors, la qual cosa és positiva. A la taula que mostrem, hem plasmat els errors de les categories general, però a l'Annex 13 mostrarem tota la taula detallada amb els noms de les categories i les subcategories, per tal de poder observar en quines subcategories es concentren la majoria d'errors.

Category	Critical	Major	Minor	Preferential
Meaning	33	31	27	1
Language	8	21	51	0
Terminology	2	4	11	0
Style	3	28	67	8
Layout	0	8	0	0
Country Standards	0	0	1	0

Repeat	1	2	10	1
Kudos	0	0	0	0
Other	0	2	1	2

Taula 7. Resultats de la classificació d'errors

Seria molt interessant mostrar també reculls d'errors per veure les correccions que ens han ofert els col·laboradors, però per qüestions de confidencialitat, això no serà possible. Tot i així, podem afirmar que gran part dels problemes més greus (*critical*) es deuen a tres factors principals: traduccions sense cap sentit (*mistranslation*), traduccions literals (*literal translation*) i text sense traduir (*untranslated text*). Aquests tres tipus d'errors són els característics de molts traductors automàtics i la principal raó per la qual molts usuaris no acudeixen als seus serveis. Considerem que són aspectes que es poden millorar amb entrenament de corpus més genèrics que ajudin a establir un llenguatge natural pel nostre motor. Val a dir també que el fet que la terminologia presenti pocs errors és una bona senyal, ja que confirmem que amb l'entrenament d'un motor amb corpus dels clients d'una empresa, el motor serà completament capaç de traduir els seus termes correctament i, per tant, poden millorar molt la productivitat del traductor.

Quant a l'avaluació amb el **rànquing de traductors automàtics**, com hem dit a l'apartat anterior, comptàvem amb la presència de 12 traductors i posteditors. De tots ells, però, 8 han acabat participant a l'avaluació. Tot i així, considerem els resultats bastant significatius i ens poden ajudar a treure bones conclusions. Quant als resultats, com era d'esperar, el nostre traductor no ha passat per davant dels altres dos. De fet, és DeepL qui ha obtingut la màxima puntuació. És normal que el nostre motor no hagi obtingut la màxima puntuació per una simple raó: hem comparat traduccions d'un text que no pertany al 100% al contingut amb què hem entrenat el traductor, és a dir, el nostre traductor ha estat entrenat amb el vocabulari i l'estil dels clients de l'empresa amb qui hem col·laborat. Si haguéssim fet l'avaluació amb un text d'algun d'aquests clients, és possible que els resultats fossin millors que els que hem obtingut ara, o potser no. Per qüestions de confidencialitat, però, i degut al fet que els posteditors que han col·laborat a l'avaluació són externs a l'empresa, no hem pogut fer l'avaluació amb un text d'aquestes característiques.

Observem, a continuació, el diagrama que mostra els resultats generals de la classificació dels tres traductors automàtics:

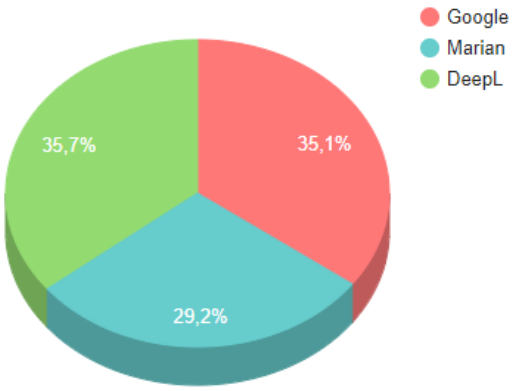


Figura 10. Diagrama de resultats del rànquing de motors

Com veiem finalment, la diferència de puntuació entre Google i DeepL ha estat mínima, però DeepL ha passat per davant dels altres dos. El nostre motor, al qual hem anomenat Marian, queda en últim lloc, però val a dir que tampoc ha tingut una puntuació tan inferior en comparació als altres. En general, els tres han quedat bastant igualats. Seguidament, observem dos diagrames que explicarem a posteriori.

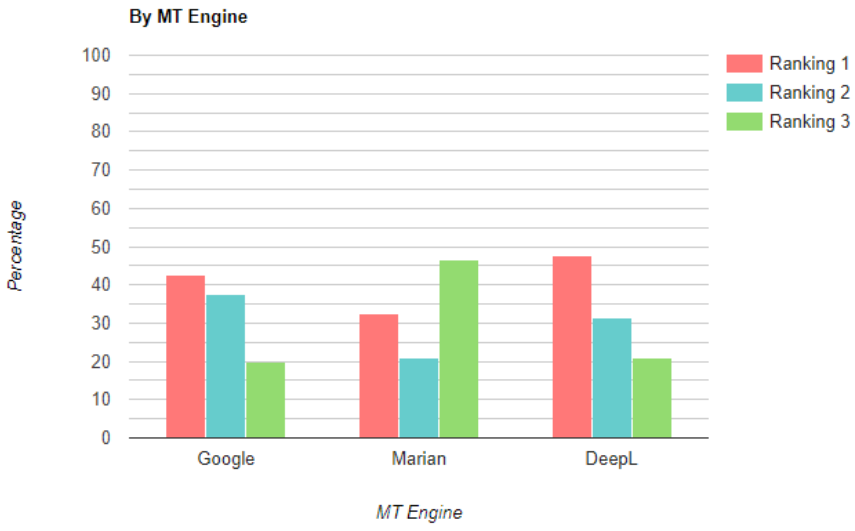


Figura 11. Resultats per motor del rànquing de motors

En aquest primer diagrama, observem els resultats que ha obtingut cada motor. Com bé hem comentat abans, els avaluadors havien de puntuar amb un 1, un 2 o un 3 cada motor en cada segment. Comprovem al diagrama que DeepL és qui ha obtingut més vegades la puntuació més alta (1), mentre que Marian és qui ha obtingut més vegades la més baixa (3). Google en general on ha obtingut més punts és a les dues primeres posicions, però tot i així no ha aconseguit atrapar a DeepL. Observem a continuació, la taula que forma aquest diagrama per tal de poder veure amb números aquesta informació:

Motor	**1**	**2**	**3**
Google Translate	42,75%	37,37%	19,87%
Marian	32,37%	21%	46,62%
DeepL	47,75%	31,25%	21%

Taula 8. Resultat per motor del rànquing de motors

A la taula comprovem que, malgrat Google és qui ha obtingut menys puntuació de 3, el fet que DeepL el superi amb la puntuació de 1 fa que li passi per davant.

El mateix podem dir del següent diagrama, on veiem una informació similar a l'anterior, però representada de diferent manera. En aquest cas, veiem per cada número de classificació, quina és la quantitat de "vots" que ha tingut cada motor. El més significatiu o impactant d'aquest rànquing, és que Marian destaca considerablement com a motor amb més vots negatius, la qual cosa és una mala senyal i un mal resultat.

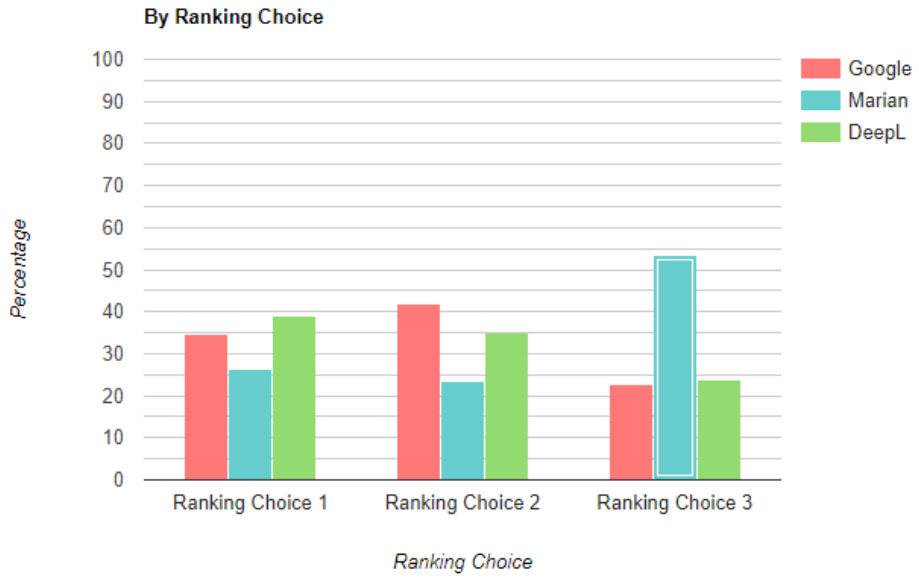


Figura 12. Resultats per puntuació del rànquing de motors

Considerem que l'últim que ens queda a dir sobre aquest rànquing és que hem de tenir en compte que la comparació es feia amb dos gegants de la indústria de la traducció automàtica. Dues empreses que compten amb milions i milions de segments alineats per entrenar els seus traductors i que compten amb professionals que es dediquen constantment a la seva evolució i els propis usuaris que col·laboren en la millora del motor. En el nostre cas, estem parlant d'un traductor automàtic que tot just ha estat creat i és la primera vegada que les seves traduccions surten a la llum. Per tant, això ens indica que queda un procés de millora del traductor força important, però que el camí que hem construït fins ara és correcte.

Com a tancament d'aquest apartat, és imprescindible mencionar que aquest procés s'haurà d'anar repetint a mesura que seguim entrenant i millorant el motor, atès que sempre serà necessària la presència dels posteditors que ens ajudin a detectar quines flaqueses té el traductor automàtic i quins aspectes haurem d'anar millorant a la llarga.

6. Conclusions

Després de l'elaboració d'aquest treball, es poden arribar a moltes conclusions. Aquestes conclusions les podem valorar de diverses maneres, ja sigui de manera quantitativa o qualitativa. Sense tenir en compte els resultats que hem obtingut, l'aprenentatge que es treu d'un projecte com aquest és molt elevat i, per un traductor, permet anar més enllà de les seves tasques i poder experimentar la part més tecnològica i científica del seu camp de treball.

Per tal de plasmar les nostres conclusions, creiem que és necessari recuperar les preguntes que ens hem plantejat a l'apartat 2. Objectius: preguntes de recerca i hipòtesi.

Pel que fa la primera pregunta i tenint en compte que es tracta d'un procés lent que requereix força temps de desenvolupar, no hem pogut arribar a comprovar si la productivitat del sector de l'empresa on s'instaura aquest traductor ha millorat o no, però vistos els estudis realitzats per altres autors i els resultats que ofereix el traductor que hem entrenat, ens mostrem positius de cara a una futura millora de la productivitat.

Quant a la segona pregunta plantejada, considero que vistos els resultats de l'avaluació quantitativa, podem dir que sí que val la pena invertir recursos en entrenar un motor personalitzat, ja que els resultats són, si més no, millors que els d'alguns genèrics. Si tenim en compte l'avaluació qualitativa, que és la que ens ha indicat uns punts claus per la millora i ens ha plasmat a la perfecció el rendiment del nostre motor, considero important plantejar futurs entrenaments incorporant més corpus, ja siguin d'àmbit genèric o més específics de l'àmbit treballat; però és important millorar la qualitat de les traduccions quant al llenguatge i l'expressió, que és on més falla.

Això últim que hem comentat ens respon a la nostra última pregunta, per la qual cosa podem dir que sí, un traductor automàtic entrenat amb corpus interns i externs centrats en un àmbit concret d'especialitat tenen uns resultats millors que alguns dels traductors automàtics més importants del mercat actual. Amb certs matisos, és clar. Com hem comentat a l'apartat anterior, quant a terminologia específica el traductor és força potent, la qual cosa en molts moments el farà passar per davant a la qualitat terminològica d'un motor genèric del mercat. Això sí, cal aprofundir en la millora de la qualitat de

l'expressió, ja que en aquest aspecte són els traductors del mercat els qui passen per davant.

Quant al futur del traductor dins l'empresa, considero que amb les millores que s'aniran implementant i amb el creixement del motor a mesura que tinguem més dades d'entrenament, facilitaran la incorporació del motor en el flux de treball. També és cert, però, que considero necessària la redacció d'una guia de postedició on es destaquin els principals errors que pot produir la màquina i facilitar i agilitzar, així, les tasques de postedició. És evident, doncs, que una empresa que pretén incorporar un traductor automàtic elaborat per ells mateixos, requereixen de recursos ja no només per entrenar-lo i posar-lo en marxa, sinó també per mantenir-lo i millorar-lo a mesura que passa el temps. Si no es duu a terme aquesta última tasca, el motor anirà quedant obsolet a mesura que el llenguatge usat en els textos variï o augmenti.

Així doncs, a mode de conclusió final, considero necessari recalcar que les tecnologies de la traducció avancen cada dia, i cada vegada les màquines evolucionaran més i més en aquest aspecte. Per això, és imprescindible que les empreses del sector es mantinguin actualitzades i vagin innovant a mesura que el temps i les tecnologies avancen, per tal de no quedar-se enrere respecte altres empreses o respecte el sector en general. Les tecnologies formaran part de la feina dels traductors i els facilitaran la feina, per això ens hem de mostrar oberts i acceptar-les com una eina de treball més.

7. Bibliografia

Arevalillo, J. J. (2012). *La traducción automática en las empresas de traducción.* Revista Tradumàtica: Tecnologies de la traducció. Desembre 2012. ISSN: 1578-7559.

Babych, B. (2014) Mètriques d'avaluació automatitzada de TA i les seves limitacions. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*. Núm. 12, pp. 464-70. Disponible a: https://www.raco.cat/index.php/Tradumatica/article/view/n12-babych Consultat el: 16/04/2020

Casacuberta i Peris (2017) Traducción automática neuronal. Revista Tradumàtica. Tecnologies de la traducción, 15, 66-74. Disponible a: https://ddd.uab.cat/pub/tradumatica/tradumatica_a2017n15/tradumatica_a2017n15p66.pdf Consultat el: 30/01/2020

Chung, T i Gilde, D. (2009) Unsupervised Tokenization for Machine Translation. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, p. 718–726, Singapore. Disponible a: https://dl.acm.org/doi/10.5555/1699571.1699606 Consultat el: 01/05/2020.

Forcada, M. (2017) Making sense of neural machine translation. Translation Spaces. 6, 291-309. Disponible a: https://www.researchgate.net/publication/321495189_Making_sense_of_neural_machine_translation Consultat el: 30/01/2020.

Guzmán, F., Abdelali, A., Temnikova, I, et al. (2015). How do Humans Evaluate Machine Translation. Disponible a: https://www.aclweb.org/anthology/W15-3059.pdf Consultat el: 16/04/2020

Hearne, M. & Way, A. (2011). Statistical Machine Translation: A Guide for Linguists and Translators. Language and Linguistics Compass, 5:5, p. 205–226.

Hutchins, W. J. & Somers, H. L. (1992). An Introduction to Machine Translation, Londres: Academic Press.

Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zenz, R., Dyer, C., Bojar, O., Constantin, A. I Herbst, E. (2007) Moses: Open Source Toolkit for Statistical Machine Translation, *Annual Meeting of the Association for Computational Linguistics (ACL)* Disponible a: https://www.aclweb.org/anthology/P07-2045.pdf Consultat el: 08/02/2020

Kudo, Taku (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. P*roceedings of the 56th Annual Meeting on Association for Computational Linguistics* - Volum 1, pp. 66-75. Disponible a: https://www.aclweb.org/anthology/P18-1007/ Consultat el: 19/05/2020

Lita, L.V., Ittycheriah, A., et al. (2003). Truecasing. P*roceedings of the 41st Annual Meeting on Association for Computational Linguistics* - Volum 1, pp. 152 - 159. (ACL 2003). Disponible a: https://doi.org/10.3115/1075096.1075116 Consultat el: 19/05/2020

Llitjós-Font, A., Carbonell, J.G., Lavie, A. (2005). A framework for interactive and automatic refinement of transfer-based machine translation. *Proceedings of the 10th Annual Conference of the European Association for Machine Translation*. Disponible a: https://kilthub.cmu.edu/articles/A_Framework_for_Interactive_and_Automatic_Refinement_of_Transfer-based_Machine_Translation/6620648/1 Consultat el: 16/04/2020

Oliver, A (2014). Traducció i tecnologies: eines, processos i recursos. Universitat Oberta de Catalunya (UOC).

Oliver, A. (2020a). MTUOC: integración de traducción automática estadística y neuronal en entornos profesionales de traducción. Universitat Oberta de Catalunya (UOC). Disponible a: https://xwiki.recursos.uoc.edu/wiki/mat00001ca/view/MTUOC%3A%20traducci%C3%B3n%20autom%C3%A1tica%20neuronal/ Consultat el: 29/04/2020.

Oliver, A. (2020b). 3. Traducció automàtica neuronal con Marian. Universitat Oberta de Catalunya (UOC). Disponible a: https://xwiki.recursos.uoc.edu/wiki/mat00001ca/view/1.4.%20Traducci%C3%B3n%20autom%C3%A1tica%20y%20postedici%C3%B3n/3.%20Traducci%C3%B3n%20autom%C3%A1tica%20neuronal%20con%20Marian/ Consultat el: 21/05/2020.

Panić, M (2019). DQF-MQM: Beyond Automatic MT Quality Metrics. TAUS. Disponible a: https://blog.taus.net/dqf-mqm-beyond-automatic-mt-quality-metrics Consultat el: 16/04/2020

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *40th Annual meeting of the Association for Computational Linguistics.* pp. 311 – 318. Disponible a: https://dl.acm.org/action/doSearch?AllField=wer Consultat el: 16/04/2020

Parra-Escartín, P. (2019) ¿Cómo ha evolucionado la traducción automática en los últimos años? Tecnología aplicada a la traducción. *La linterna del traductor*. Disponible a: http://www.lalinternadeltraductor.org/n16/traduccion-automatica.html Consultat el: 09/03/2020.

Popovic, Maja (2018). Error Classification and Analysis for Machine Translation Quality Assessment. Disponible a: file:///C:/Users/mmend/Downloads/err-class.pdf Consultat el: 16/04/2020

Rico. C. i A. Martín de Santa Olalla (1997): "New Trends in Machine Translation", *Meta*, vol 4, nº4, 605-615. Disponible a: https://www.erudit.org/en/journals/meta/1997-v42-n4-meta173/003822ar/ Consultat el: 25/02/2020.

Sennrich, R. et al. (2016). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL 2016). Disponible a: https://arxiv.org/abs/1508.07909 Consultat el: 19/05/2020

Snover et al. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the 7th Conference of the Association for Machine Translation in

the Americas, pages 223-231. Cambridge. Disponible a: http://mt-archive.info/AMTA-2006-Snover.pdf Consultat el: 14/04/2020

Tomás, J., Más, J., Casacuberta, F. (2003). A Quantitative Method for Machine Translation Evaluation. Association for Computational Linguistics. Disponible en: https://www.aclweb.org/anthology/W03-2804 Consultat el: 08/04/2020

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. Disponible a: https://ddd.uab.cat/record/166753. Consultat el: 05/03/2020.

Trujillo, A. (1999) *Translation Engines: Techniques for Machine Translation.* Springer.

Vilar, D., Leusch, G., Ney, H. et al. (2007) Human Evaluation of machine translation through binary system comparisons. *Proceedings of the Second Workshop on Statistical Machine Translation,* pp. 96–103. Association for Computational Linguistics. Disponible a: https://www.aclweb.org/anthology/W07-0713.pdf Consultat el: 16/04/2020

Vilar, D., Xu, J., D'Haro, LF, et al. (2006) Error Analysis of Statistical Machine Translation Output. *Proceedings of 5th International Conference on Language Resources and Evaluation,* pp. 697–702. Disponible a: http://www.lrec-conf.org/proceedings/lrec2006/pdf/413_pdf.pdf Consultat el: 16/04/2020

Villegas, M., Intxaurrondo, A., Gonzalez-Agirre, A., et al. (2018). *The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations.* In LREC MultilingualBIO: Multilingual Biomedical Text Processing. ELRA..

White, J.S., O'Connell, T.A., O'Mara, F.E. (1994) The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the workshop on Human Language Technology*. Disponible a: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.1288&rep=rep1&type=pdf Consultat el: 16/04/2020

Zaretskaya (2017) Machine Translation Post-Editing at TransPerfect – the “Human” Side of the Process.

Zhang, Y., Vogel, S., Waibel, A. (2004). Interpreting BLEU/NIST Scores: How Much Improvement Do We Need to Have a Better System? Language Resources and Evaluation Conference 2004. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.536.5232&rep=rep1&type=pdf Consultat el: 08/04/2020.

Annexos

Annex 1

Llistat de traductors automàtics i sistemes d'entrenament de TA del mercat

Google Translate (Auto ML)	Microsoft Translator Hub	ITranslate 4	Globalese MT	Amagama
Reverso	PROMT Translator	Wordlingo	SAP Translation Hub	Lilt
DeepL	Moses	OpenTrad	Amazon translate	CapitaMT
Yandex	MTradumàtica	Baidu API machine translation	TildeMT	Weblate
Apertium	KantanMT	Glosbe	Nematus	OpenLogos
Lucy Software	ModenMT	Netease Sight API machine translation	Tensorflow	
Systran	Slate Desktop Pro	Youdao Zhiyun API machine translation	MTUOC	

Annex 2

Taula de motors i característiques

Motor	Ubicació servidors	Volum	Cost	Altres serveis	API? Personalització?	Format arxius	Llengües
Google Translator Hub	Google	Mínim 100 per validation i 100 per test + els d'entrenament. Màxim 15 milions de parells.	Preus d'entrenament per hora i de traducció per caràcters.	-	Sí / Sí	.tsv, .csv, .tmx	103 llengües diferents
DeepL	Propi de DeepL (alt grau de privacitat)	-	Ultimate 34€/usuari/mes		Sí/No	-	72 combinacions
Apertium	-	-	Gratuït (opensource)	-	Sí/No	-	30 combinacions
Lucy Software	Personal	-	Gratuït (opensource)	-	Sí/Sí per part de l'empresa	-	36 combinacions

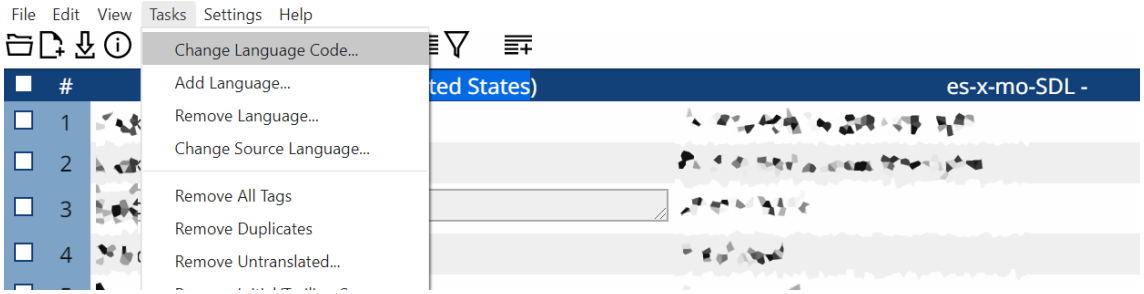
PROMT Translator	Personal	Il·limitat	750€	-	Sí/Sí	-	14 combinacions
Moses	Personal	Il·limitat	Gratuït (opensource)	-	Sí/Sí (creació motor)	.tmx, .txt	Qualsevol
MTradumàtica	Personal	Il·limitat	Gratuït (opensourse, Moses)	QA amb mètriques BLEU, etc.	Sí/Sí (creació motor)	TMX, etc. (tots)	Qualsevol
KantanMT[2]	In top-tier data centers.	Il·limitat	Contactar per demanar preus	-	Sí / Sí (creació motor)	Tots	750 combinacions lingüístiques
Microsoft Translator	Microsoft	-	Diferents tarifes depenent d'allò que busques		Sí/Sí (creació motor)	.tmx	60 llengües +
Amazon Translate	Amazon	-	Uns 14€ per milió de caràcters	-	Sí/Sí, només personalització de terminologia	.tmx i .csv	32 idiomes

Annex 3

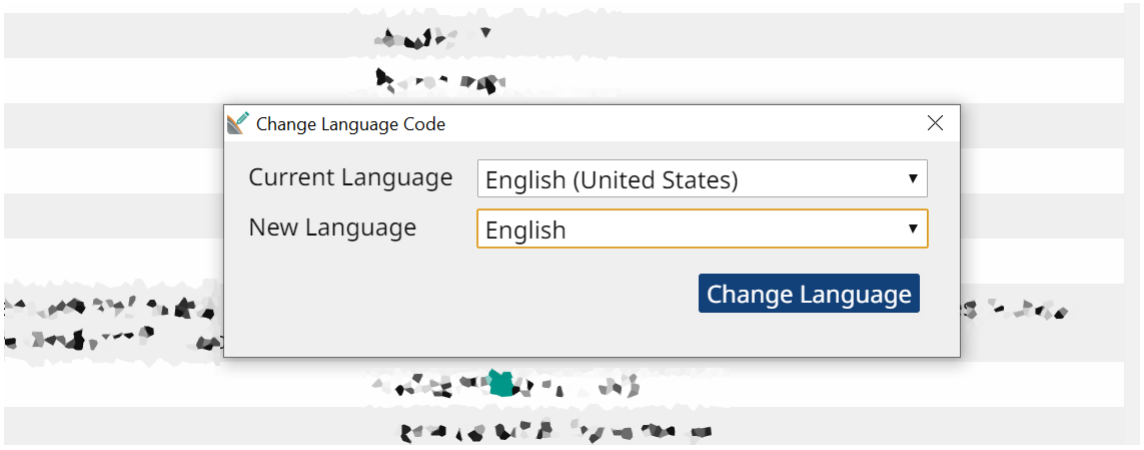
TMXEditor

Obrim el programa, anem a File > Open i triem l'arxiu .tmx que volem editar. Un cop l'obrim, la vista que tenim és la següent:

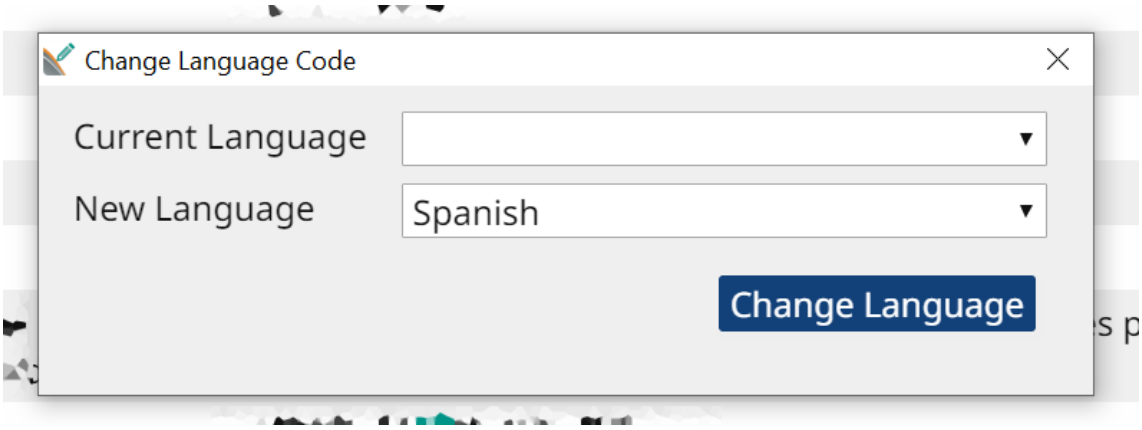
Com veiem, les llengües que hi ha indicades no són l'anglès i l'espanyol senzill que volem posar, sinó que l'anglès té la varietat d'Estats Units configurada i l'espanyol té el codi de SDL Trados. Per tal de poder canviar els codis, anem a Tasks > Change Language Code...:



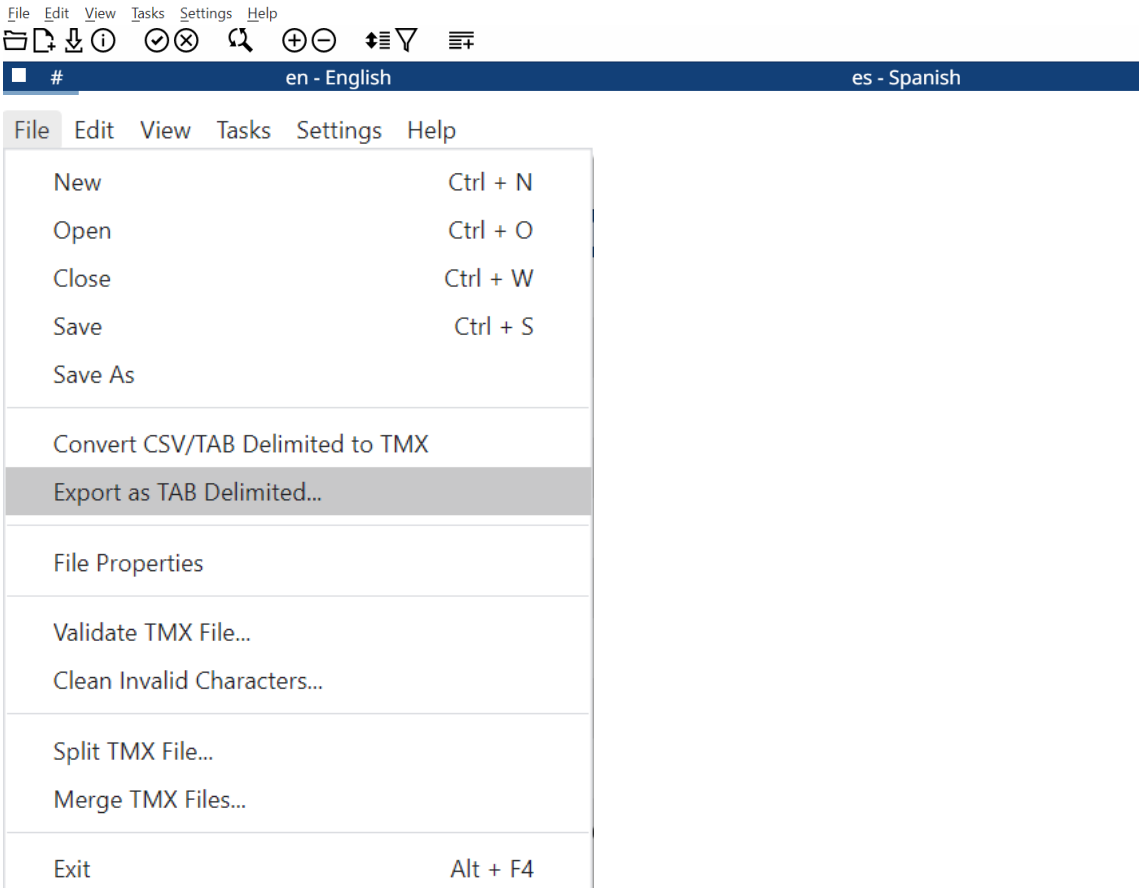
Canviem la llengua original a l'anglès sense varietat (English):



Com veurem a continuació, el programa no detecta la segona llengua, ja que és una varietat que no té incorporada, tot i així ens permet seleccionar-la i canviar-la a espanyol sense varietat (Spanish):



Un cop ho hem canviat, ja podem exportar la memòria a format text pla tabulat, anant a File > Export as TAB Delimited...:



A partir d'aquí, li posem a l'arxiu el nom que vulguem i el guardem. El programa també ens permet fer altres accions com eliminar etiquetes, segments que no s'han traduït, segments duplicats, etc. Podem dur a terme aquestes accions però hem de tenir present que no sempre són del tot fiables i ens poden quedar etiquetes que malmetran l'entrenament.

Annex 4

Tokenització

```
/Exemples$ python3 MTUOC_tokenizer_eng.py < corpus_en > corpus.tok.en
/Exemples$ python3 MTUOC_tokenizer_spa.py < corpus_es > corpus.tok.es
```

Un cop passem la indicació pel terminal, obtenim els dos arxius desitjats "corpus.tok.en" i "corpus.tok.es". A continuació, mostrem el canvi que s'observa al text.

```
66 Put the tablet into the barrel Put the plunger back onto the syringe For the 15 mg tablet:
67 Therefore, the consolidation of the professional master's degree programs is a challenge to
   be overcome .
68 Sodium oxybate has a well known abuse potential.
69 Renal and urinary disorders: common: rare: very rare:
70 Lornaxicam when used as adjunct to lignocaine provided better analgesia compared to the
   control.
71 Enough representation of patients with this diagnosis was not made, and is not a reflection
   on the incidence or prevalence of this pathology in a national level.
72 Burning sensation and pruritus were very common, usually mild to moderate in severity and
   tended to resolve within one week of starting treatment.
73 Use in patients with hepatic impairment
74 Confusion, sleep disorder, libido decreased Suicide ideation Suicide, suicide attempts,
   aggressive behaviour (sometimes directed against others), psychosis including hallucinations
75 Other risk factors associated with PHN include severe pain during HZ, more severe skin
   exanthema, greater nerve compromise in the affected dermatome, more severe immune response,
   and psychological and social factors.
76 This study had a low risk of bias 89 level of evidence: high.
```

En aquesta primera imatge, observem els segments del 66 al 76 del corpus "corpus_en" sense ser tokenitzat, on podem comprovar que tots els signes de puntuació es troben col·locats de manera normal.

```
66 Put the tablet into the barrel Put the plunger back onto the syringe For the 15 mg tablet :
67 Therefore , the consolidation of the professional master 's degree programs is a challenge
   to be overcome .
68 Sodium oxybate has a well known abuse potential .
69 Renal and urinary disorders : common : rare : very rare :
70 Lornaxicam when used as adjunct to lignocaine provided better analgesia compared to the
   control .
71 Enough representation of patients with this diagnosis was not made , and is not a reflection
   on the incidence or prevalence of this pathology in a national level .
72 Burning sensation and pruritus were very common , usually mild to moderate in severity and
   tended to resolve within one week of starting treatment .
73 Use in patients with hepatic impairment
74 Confusion , sleep disorder , libido decreased Suicide ideation Suicide , suicide attempts ,
   aggressive behaviour ( sometimes directed against others ) , psychosis including
   hallucinations
75 Other risk factors associated with PHN include severe pain during HZ , more severe skin
   exanthema , greater nerve compromise in the affected dermatome , more severe immune
   response , and psychological and social factors .
76 This study had a low risk of bias 89 level of evidence : high .
```

En canvi, en aquesta segona imatge, observem els mateixos segments del corpus "corpus.tok.en", on comprovem que ara sí que els signes de puntuació es troben separats de les paraules, és a dir, s'ha separat el corpus per tokens.

Podem comprovar també que ha passat el mateix en el corpus en espanyol. De la mateixa manera que a l'exemple anterior, a la primera imatge mostrarem els segments originals en espanyol i a la segona els segments tokenitzats.

```
66 Para el comprimido de 15 mg: llene la jeringa con 4 ml de agua.
67 Así, la consolidación de los programas de maestría profesional representa un reto a ser
   vencido .
68 El oxibato de sodio tiene un potencial de abuso bien conocido.
69 Trastornos renales y urinarios: frecuentes: raros: muy raros:
70 El lornaxicam usado como adyuvante de lidocaína suministró una mejor analgesia en
   comparación con el grupo control.
71 No puede ser lo suficientemente representativa de pacientes que con este diagnóstico se
   atienden en todos los hospitales costarricenses, ni es reflejo de incidencia o prevalencia
   de esta patología en el nivel nacional.
72 La sensación de quemazón y prurito fueron muy frecuentes, habitualmente de intensidad leve o
   moderada, y tendieron a resolverse una semana después de iniciarse el tratamiento.
73 Uso en pacientes con insuficiencia hepática
74 Confusión, alteración del sueño, disminución de la libido Ideación suicida Suicidio,
   intentos de suicidio, comportamiento agresivo (a veces hacia otras personas), psicosis
   incluyendo
75 Otros factores de riesgo que se han asociado con la NPH son: dolor severo durante el HZ,
   mayor severidad del exantema cutáneo, mayor compromiso neurológico en el dermatoma afectado,
   una respuesta inmune más severa y factores psicosociales.
76 Este estudio tiene bajo riesgo de sesgos 89.
```

```
66 Para el comprimido de 15 mg : llene la jeringa con 4 ml de agua .
67 Así , la consolidación de los programas de maestría profesional representa un reto a ser
   vencido .
68 El oxibato de sodio tiene un potencial de abuso bien conocido .
69 Trastornos renales y urinarios : frecuentes : raros : muy raros :
70 El lornaxicam usado como adyuvante de lidocaína suministró una mejor analgesia en
   comparación con el grupo control .
71 No puede ser lo suficientemente representativa de pacientes que con este diagnóstico se
   atienden en todos los hospitales costarricenses , ni es reflejo de incidencia o prevalencia
   de esta patología en el nivel nacional .
72 La sensación de quemazón y prurito fueron muy frecuentes , habitualmente de intensidad leve
   o moderada , y tendieron a resolverse una semana después de iniciarse el tratamiento .
73 Uso en pacientes con insuficiencia hepática
74 Confusión , alteración del sueño , disminución de la libido Ideación suicida Suicidio ,
   intentos de suicidio , comportamiento agresivo ( a veces hacia otras personas ) , psicosis
   incluyendo
75 Otros factores de riesgo que se han asociado con la NPH son : dolor severo durante el HZ ,
   mayor severidad del exantema cutáneo , mayor compromiso neurológico en el dermatoma
   afectado , una respuesta inmune más severa y factores psicosociales .
76 Este estudio tiene bajo riesgo de sesgos 89 .
```

Annex 5

Truecasing

Com hem dit a l'explicació de l'apartat 4.5.2. Truecasing, primerament hem de passar les indicacions per crear els models de truecasing d'ambdues llengües, com veiem a la figura següent:

```
marta@marta-HP-Pavilion-x360-Convertible-14-dh1xxx:~/Escritorio/Exemples$ python3 MTUOC_train_truecaser.py -n tc.en -c corpus.tok.en -d eng.dict
Reading corpus: corpus.tok.en
Reading dictionary: eng.dict
marta@marta-HP-Pavilion-x360-Convertible-14-dh1xxx:~/Escritorio/Exemples$ python3 MTUOC_train_truecaser.py -n tc.es -c corpus.tok.es -d spa.dict
Reading corpus: corpus.tok.es
Reading dictionary: spa.dict
```

Un cop hem entrenat els models, com ja hem comentat, ja podem dur a terme el procés de truecasing, tal i com veiem a la figura:

```
o/Exemples$ python3 MTUOC_truecaser.py tc.en < corpus.tok.en > corpus.true.en
o/Exemples$ python3 MTUOC_truecaser.py tc.es < corpus.tok.es > corpus.true.es
```

A continuació, mostrem els resultats fent servir els mateixos segments que hem vist a l'annex anterior. Primerament, observem l'arxiu tokenitzat en anglès:

```
66 Put the tablet into the barrel Put the plunger back onto the syringe For the 15 mg tablet :
67 Therefore , the consolidation of the professional master 's degree programs is a challenge
   to be overcome .
68 Sodium oxybate has a well known abuse potential .
69 Renal and urinary disorders : common : rare : very rare :
70 Lornaxicam when used as adjunct to lignocaine provided better analgesia compared to the
   control .
71 Enough representation of patients with this diagnosis was not made , and is not a reflection
   on the incidence or prevalence of this pathology in a national level .
72 Burning sensation and pruritus were very common , usually mild to moderate in severity and
   tended to resolve within one week of starting treatment .
73 Use in patients with hepatic impairment
74 Confusion , sleep disorder , libido decreased Suicide ideation Suicide , suicide attempts ,
   aggressive behaviour ( sometimes directed against others ) , psychosis including
   hallucinations
75 Other risk factors associated with PHN include severe pain during HZ , more severe skin
   exanthema , greater nerve compromise in the affected dermatome , more severe immune
   response , and psychological and social factors .
76 This study had a low risk of bias 89 level of evidence : high .
```

I el resultat del procés de truecasing pels mateixos segments en anglès:

```
66 put the tablet into the barrel put the plunger back onto the syringe for the 15 mg tablet :
67 therefore , the consolidation of the professional master 's degree programs is a challenge
   to be overcome .
68 sodium oxybate has a well known abuse potential .
69 renal and urinary disorders : common : rare : very rare :
70 Lornaxicam when used as adjunct to lignocaine provided better analgesia compared to the
   control .
71 enough representation of patients with this diagnosis was not made , and is not a reflection
   on the incidence or prevalence of this pathology in a national level .
72 burning sensation and pruritus were very common , usually mild to moderate in severity and
   tended to resolve within one week of starting treatment .
73 use in patients with hepatic impairment
74 confusion , sleep disorder , libido decreased suicide ideation suicide , suicide attempts ,
   aggressive behaviour ( sometimes directed against others ) , psychosis including
   hallucinations
75 other risk factors associated with PHN include severe pain during Hz , more severe skin
   exanthema , greater nerve compromise in the affected dermatome , more severe immune
   response , and psychological and social factors .
76 this study had a low risk of bias 89 level of evidence : high .
```

El mateix ha passat amb els exemples en espanyol. Observem primer l'arxiu tokenitzat en espanyol:

```
66 Para el comprimido de 15 mg : llene la jeringa con 4 ml de agua .
67 Así , la consolidación de los programas de maestría profesional representa un reto a ser
   vencido .
68 El oxibato de sodio tiene un potencial de abuso bien conocido .
69 Trastornos renales y urinarios : frecuentes : raros : muy raros :
70 El lornaxicam usado como adyuvante de lidocaína suministró una mejor analgesia en
   comparación con el grupo control .
71 No puede ser lo suficientemente representativa de pacientes que con este diagnóstico se
   atienden en todos los hospitales costarricenses , ni es reflejo de incidencia o prevalencia
   de esta patología en el nivel nacional .
72 La sensación de quemazón y prurito fueron muy frecuentes , habitualmente de intensidad leve
   o moderada , y tendieron a resolverse una semana después de iniciarse el tratamiento .
73 Uso en pacientes con insuficiencia hepática
74 Confusión , alteración del sueño , disminución de la libido Ideación suicida Suicidio ,
   intentos de suicidio , comportamiento agresivo ( a veces hacia otras personas ) , psicosis
   incluyendo
75 Otros factores de riesgo que se han asociado con la NPH son : dolor severo durante el HZ ,
   mayor severidad del exantema cutáneo , mayor compromiso neurológico en el dermatoma
   afectado , una respuesta inmune más severa y factores psicosociales .
76 Este estudio tiene bajo riesgo de sesgos 89 .
```

I el resultat del procés de truecasing d'aquest arxiu en espanyol:

```
66 para el comprimido de 15 mg : llene la jeringa con 4 ml de agua .
67 así , la consolidación de los programas de maestría profesional representa un reto a ser
   vencido .
68 el oxibato de sodio tiene un potencial de abuso bien conocido .
69 trastornos renales y urinarios : frecuentes : raros : muy raros :
70 el lornaxicam usado como adyuvante de lidocaína suministró una mejor analgesia en
   comparación con el grupo control .
71 no puede ser lo suficientemente representativa de pacientes que con este diagnóstico se
   atienden en todos los hospitales costarricenses , ni es reflejo de incidencia o prevalencia
   de esta patología en el nivel nacional .
72 la sensación de quemazón y prurito fueron muy frecuentes , habitualmente de intensidad leve
   o moderada , y tendieron a resolverse una semana después de iniciarse el tratamiento .
73 uso en pacientes con insuficiencia hepática
74 confusión , alteración del sueño , disminución de la libido ideación suicida suicidio ,
   intentos de suicidio , comportamiento agresivo ( a veces hacia otras personas ) , psicosis
   incluyendo
75 otros factores de riesgo que se han asociado con la NPH son : dolor severo durante el Hz ,
   mayor severidad del exantema cutáneo , mayor compromiso neurológico en el dermatoma
   afectado , una respuesta inmune más severa y factores psicosociales .
76 este estudio tiene bajo riesgo de sesgos 89 .
```

Podem observar també en aquest exemples, el fenomen comentat a l'apartat 4.5.2. Truecasing en què la paraula "Lornaxicam" ha mantingut la majúscula en el corpus anglès, però l'ha eliminat en l'espanyol.

Annex 6

Neteja de segments llargs

Observem la indicació mostrada a l'apartat 4.5.3. Neteja de segments llargs, en el nostre terminal.

```
les$ python3 MTUOC_cleaning.py corpus.tok en es 80
```

Podem comprovar que ha funcionat obrint els arxius dels corpus abans i després de ser netejats, i observar si el nombre de línies ha disminuït o no. També cabria la possibilitat que no disminuís, donat el fet el corpus només contingui oracions curtes o de menys tokens que els que hem marcat.

Observem doncs, primerament, els arxius "corpus.tok.en" i "corpus.tok.es", que són els arxius del corpus abans de ser netejat dels segments llargs:

```
783035 This paradigm change movement is evident in nursing .
```

```
783035 Ese movimiento de cambio paradigmático es evidente en la enfermería .
```

En aquestes imatges veiem que l'última línia dels dos corpus és la número 783035. Anem a comprovar els arxius de sortida, "corpus.tok.clean.en" i "corpus.tok.clean.es":

```
765907 This paradigm change movement is evident in nursing .
```

```
765907 Ese movimiento de cambio paradigmático es evidente en la enfermería .
```

En aquest cas, observem que l'última línia dels arxius és la 765.907 i, per tant, podem confirmar que el procés ha donat resultats.

Annex 7

Tractament d'expressions numèriques

Com ja s'ha explicat a l'apartat 4.5.4. Tractament d'expressions numèriques, aquesta és la indicació que posem a la línia de comandes:

```
torio/Exemples$ python3 MTUOC_splitnumbers.py < corpus.true.en > corpus.split.en
torio/Exemples$ python3 MTUOC_splitnumbers.py < corpus.true.es > corpus.split.es
```

Observem el canvi entre els corpus "corpus.true.en" i "corpus.split.en". Com veiem, a la primera imatge hi tenim els números escrits de manera normal, mentre que a la segona imatge hi veiem els números separats per un espai, tal i com hem manat al programa que fes. El mateix ha passat als arxius en espanyol.

```
 2 currently we have three dozens of national nursing journals that disseminate the knowledge
   of the Brazilian nursing .
 3 how has Kuvan been studied ?
 4 the patient should undergo preoperative and surgical risk evaluation since the majority of
   patients with esophageal cancer are chronic alcoholics and smokers .
 5 it is known that there are controversies about which should be the most adequate model for
   the PPS .
 6 in the University hospital of Santa Catarina , the prevalence of MS according to the IDF
   definition was 21.9 % among women and 19.4 % among men .
 7 colonic polypectomy is the most important tool for stopping adenoma-cancer , and the inject
   and cut technique has demonstrated efficacy and safety in studies conducted in other
   countries .
 8 study 22
 9 in January 2002 , during the Eduardo Duhalde administration , and in the context of economic
   activity levels plunging 10 percentage points , the " public emergency and currency regime
   reform act , " law 25.561 , was enacted , declaring a public emergency in social , economic
   management , financial , and currency-related matters .
10 many investigations have demonstrated that fibers with aspect ratio greater than 100 usually
   reduce workability and produce non-uniform fibers distribution when applying conventional
   mixing procedures and techniques .
```

```
 2 currently we have three dozens of national nursing journals that disseminate the knowledge
   of the Brazilian nursing .
 3 how has Kuvan been studied ?
 4 the patient should undergo preoperative and surgical risk evaluation since the majority of
   patients with esophageal cancer are chronic alcoholics and smokers .
 5 it is known that there are controversies about which should be the most adequate model for
   the PPS .
 6 in the University hospital of Santa Catarina , the prevalence of MS according to the IDF
   definition was 2 1 . 9 % among women and 1 9 . 4 % among men .
 7 colonic polypectomy is the most important tool for stopping adenoma-cancer , and the inject
   and cut technique has demonstrated efficacy and safety in studies conducted in other
   countries .
 8 study 2 2
 9 in January 2 0 0 2 , during the Eduardo Duhalde administration , and in the context of
   economic activity levels plunging 1 0 percentage points , the " public emergency and
   currency regime reform act , " law 2 5 . 5 6 1 , was enacted , declaring a public emergency
   in social , economic management , financial , and currency-related matters .
10 many investigations have demonstrated that fibers with aspect ratio greater than 1 0 0
   usually reduce workability and produce non-uniform fibers distribution when applying
   conventional mixing procedures and techniques .
```

Annex 8

Càlcul de subwords

Observem, com hem comentat a l'apartat 4.5.6. Càlcul de subwords, la indicació del terminal:

```
20-spa-eng$ python3 MTUOC_BPE.py apply codes_file < corpus_en > corpus.BPE.en
```

Un cop executem aquesta indicació, se'ns crea l'arxiu "corpus.BPE.en" amb el càlcul de subwords fet, com veurem a continuació. Primerament, però, recordem com era l'aspecte del corpus normal, que recordem que en ser l'arxiu "corpus_en", no ha passat pels processos anteriors i, per tant, l'aspecte és el d'un text normal.

```
226 So many patients tell us that they have gone through great suffering and only found comfort
    in the nurse N. Thus, the nursing care is substantiated through individualized care for to
    each human being, based on human attitudes which transcend the linearity of specific actions.
227 In this perspective, a new way of seeing and performing extension demands is suggested,
    among other propositions, new patterns of relationship among professionals, academics and
    community.
228 But also, it can be the door of the liberation of men and women with critical thinking, able
    to coexist peacefully with the differences, managers of their own destiny, builders of
    societies equitable and respectful of the environment, in the end human beings more human.
229 This procedure allowed the isolated analysis of the elevated BP levels and medication use as
    potential determinants of temporary incapacity.
230 Premedication was midazolam 0.8 mg.kg.
231 This also applies to newborn babies and patients with an increased risk of thrombosis or "
    DIC", disseminated intravascular coagulation, a disease in which the blood coagulation
    system is disturbed.
232 Its ability to reposition the hindpaw and toes was evaluated.
233 The average score of the instrumental social support was 3.81 ± 0.69, indicating a good
    availability of perceived support, considering that the score ranges from 1 to 5, and the
    higher the value, the better the social support.
```

Corpus després del càlcul de subwords:

```
226 S@@ o many patients tell us that they have gone through great suffering and only found
    comfort in the nurse N. Th@@ us@@ , the nursing care is substantiated through individualized
    care for to each human be@@ ing@@ , based on human attitudes which transc@@ end the
    linearity of specific action@@ s.
227 In this perspec@@ tiv@@ e@@ , a new way of seeing and performing extension demands is sugg@@
    est@@ ed@@ , among other proposi@@ tions@@ , new patterns of relationship among profession@@
    al@@ s@@ , academics and commun@@ ity.
228 But al@@ so@@ , it can be the door of the liberation of men and women with critical thin@@
    king@@ , able to coexist peac@@ efully with the differ@@ ences@@ , managers of their own
    destin@@ y@@ , builders of societies equitable and respectful of the environment@@ , in the
    end human beings more human@@ .
229 This procedure allowed the isolated analysis of the elevated BP levels and medication use as
    potential determinants of temporary incapaci@@ ty.
230 Pre@@ medication was midazolam 0.@@ 8 mg.@@ kg.
231 This also applies to newborn babies and patients with an increased risk of thrombosis or "
    DIC@@ "@@ , disseminated intravascular coagul@@ ation@@ , a disease in which the blood
    coagulation system is disturb@@ ed.
232 I@@ ts ability to re@@ position the hind@@ paw and toes was evalu@@ ated.
233 The average score of the instrumental social support was 3.@@ 8@@ 1 ± 0.@@ 6@@ 9@@ ,
    indicating a good availability of perceived support@@ , considering that the score ranges
    from 1 to 5@@ , and the higher the valu@@ e@@ , the better the social support@@ .
```

Annex 9

Dades de validació

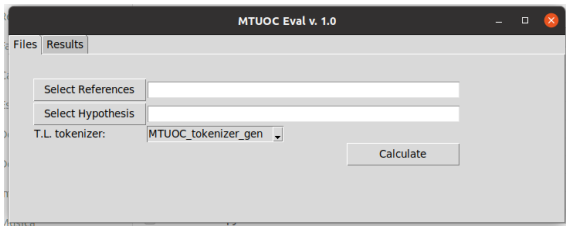
A continuació, observem les dades de validació extretes de l'arxiu vàlid.log, a partir de les quals hem elaborat el diagrama que es mostra a l'apartat 5. Resultats.

[2020-03-12 16:02:52] [valid] Ep. 1 : Up. 10000 : translation : 0 : new best
[2020-03-12 17:19:17] [valid] Ep. 1 : Up. 10000 : translation : 0.666 : new best
[2020-03-12 18:13:51] [valid] Ep. 1 : Up. 20000 : translation : 0.721 : new best
[2020-03-12 19:08:25] [valid] Ep. 1 : Up. 30000 : translation : 0.72 : stalled 1 times
[2020-03-12 20:03:07] [valid] Ep. 1 : Up. 40000 : translation : 0.735 : new best
[2020-03-12 20:57:44] [valid] Ep. 1 : Up. 50000 : translation : 0.742 : new best
[2020-03-12 21:52:58] [valid] Ep. 2 : Up. 60000 : translation : 0.752 : new best
[2020-03-12 22:47:40] [valid] Ep. 2 : Up. 70000 : translation : 0.738 : stalled 1 times
[2020-03-12 23:42:19] [valid] Ep. 2 : Up. 80000 : translation : 0.739 : stalled 2 times
[2020-03-13 00:36:55] [valid] Ep. 2 : Up. 90000 : translation : 0.741 : stalled 3 times
[2020-03-13 01:31:27] [valid] Ep. 2 : Up. 100000 : translation : 0.745 : stalled 4 times
[2020-03-13 02:26:28] [valid] Ep. 3 : Up. 110000 : translation : 0.737 : stalled 5 times
[2020-03-13 17:32:15] [valid] Ep. 1 : Up. 10000 : translation : 0.691 : new best
[2020-03-13 18:21:50] [valid] Ep. 1 : Up. 20000 : translation : 0.751 : new best
[2020-03-13 19:11:29] [valid] Ep. 1 : Up. 30000 : translation : 0.763 : new best
[2020-03-13 20:01:19] [valid] Ep. 1 : Up. 40000 : translation : 0.765 : new best
[2020-03-13 20:51:08] [valid] Ep. 1 : Up. 50000 : translation : 0.779 : new best
[2020-03-13 21:41:14] [valid] Ep. 2 : Up. 60000 : translation : 0.781 : new best
[2020-03-13 22:30:54] [valid] Ep. 2 : Up. 70000 : translation : 0.776 : stalled 1 times
[2020-03-13 23:20:34] [valid] Ep. 2 : Up. 80000 : translation : 0.779 : stalled 2 times
[2020-03-14 00:10:17] [valid] Ep. 2 : Up. 90000 : translation : 0.776 : stalled 3 times
[2020-03-14 00:59:59] [valid] Ep. 2 : Up. 100000 : translation : 0.77 : stalled 4 times
[2020-03-14 01:50:08] [valid] Ep. 3 : Up. 110000 : translation : 0.775 : stalled 5 times

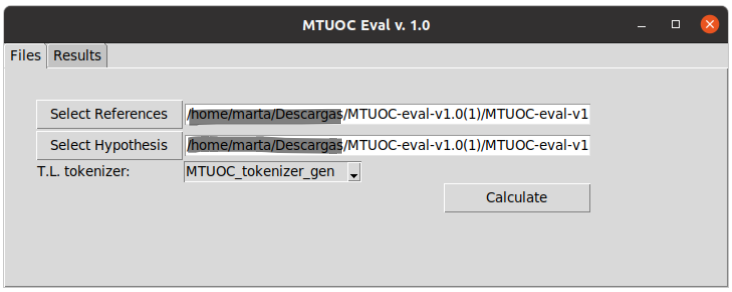
Annex 10

Com ja hem comentat a l'apartat 4.7.1 Avaluació quantitativa, disposem, a part de la línia de comandes, d'una interfície gràfica que fa més accessible l'avaluació per a aquells que ho prefereixin. A continuació, mostrem unes imatges on observem la interfície i les seves característiques.

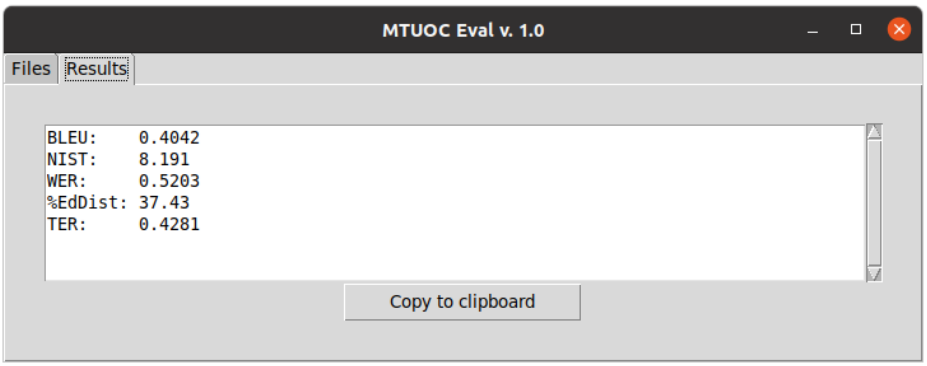
Aquest és l'aspecte de la interfície. Observem que té dues pestanyes; la primera on es gestionen els arxius que es necessiten per fer l'avaluació, i la segona on podrem observar els resultats:



Seleccionem l'arxiu de referència, l'arxiu d'hipòtesi, que és el que hem traduït amb la màquina que vulguem avaluar, i l'arxiu de tokenització de la llengua d'arribada, en el nostre cas, l'espanyol:



Seguidament, doncs, fem clic a "Calculate" i un cop el programa hagi acabat de calcular les mètriques d'avaluació, podrem accedir a la pestanya de "Results" per poder observar els resultats:



En aquest cas, els resultats són diferents que els de l'apartat on hem explicat l'avaluació que hem dut a terme perquè s'ha avaluat amb uns arxius diferents, simplement per poder observar el procés i mostrar-lo.

Annex 11

Arxiu de preparació de classificació de motors

ID	Source Segment	Segment Origin	Google	DeepL	Marian
1	Introduction: Multiple interventions are performed in critical patients admitted to Intensive Care Units (ICUs).	SampleText_1	Introducción: se realizan múltiples intervenciones en pacientes críticos ingresados en Unidades de Cuidados Intensivos (UCI).	Introducción: Se realizan múltiples intervenciones en pacientes críticos ingresados en las Unidades de Cuidados Intensivos (UCI).	Introducción: se realizan múltiples intervenciones en pacientes críticos ingresados en unidades de cuidados intensivos (UCI).
2	This study explores the presence in the daily practice of ICUs of elements related to the 6 bioethics quality indicators of the Spanish Society of Intensive and Critical Care Medicine and Coronary Units, and the participation of their members in the hospital ethics committees.	SampleText_1	Este estudio explora la presencia en la práctica diaria de las UCI de elementos relacionados con los 6 indicadores de calidad de bioética de la Sociedad Española de Medicina Intensiva y Crítica y Unidades Coronarias, y la participación de sus miembros en los comités de ética del hospital.	Este estudio explora la presencia en la práctica diaria de las UCI de elementos relacionados con los 6 indicadores de calidad bioética de la Sociedad Española de Medicina Intensiva y Crítica y de las Unidades Coronarias, y la participación de sus miembros en los comités de ética hospitalaria.	Este estudio explora la presencia en la práctica diaria de las UCI de elementos relacionados con los 6 indicadores de calidad bioética de la sociedad española de medicina intensiva y crítica y unidades coronarias, y la participación de sus miembros en los comités de ética hospitalaria.
91	In the double oral dose group, patients receive 40 mg oral esomeprazole twice daily for 11 days and followed by 40 mg once daily for 14 days.	SampleText_6	En el grupo de dosis doble oral, los pacientes reciben 40 mg de esomeprazol por vía oral dos veces al día durante 11 días y seguidos de 40 mg una vez al día durante 14 días.	En el grupo de doble dosis oral, los pacientes reciben 40 mg de esomeprazol oral dos veces al día durante 11 días y seguido de 40 mg una vez al día durante 14 días.	En el grupo de dosis doble oral, los pacientes reciben 40 mg de esomeprazol oral dos veces al día durante 11 días y seguidos por 40 mg una vez al día durante 14 días.
92	Rabies can manifest itself in two different ways: dumb rabies and furious rabies.	SampleText_6	La rabia puede manifestarse de dos maneras diferentes: rabia tonta y rabia furiosa.	La rabia puede manifestarse de dos maneras diferentes: rabia tonta y rabia furiosa.	La rabia puede manifestarse en dos formas diferentes: rabia dumb y rabia furiosa.
93	They must agree to use a reliable method of birth control during the study and for 2 months following the last dose of study drug.	SampleText_6	Deben aceptar utilizar un método anticonceptivo confiable durante el estudio y durante los 2 meses posteriores a la última dosis del medicamento del estudio.	Deben acordar usar un método fiable de control de la natalidad durante el estudio y durante los 2 meses siguientes a la última dosis del medicamento en estudio.	Deben estar de acuerdo a utilizar un método anticonceptivo confiable durante el estudio y durante 2 meses después de la última dosis del medicamento.

Annex 12

Instruccions pels posteditors (en castellà)

Les instruccions es van redactor en castellà perquè vam comptar amb la col·laboració de posteditors la llengua dels quals era el castellà i no coneixen el català. A part, per un encàrrec normal, és preferible entregar les instruccions en una de les dues llengües que es treballa en el projecte.

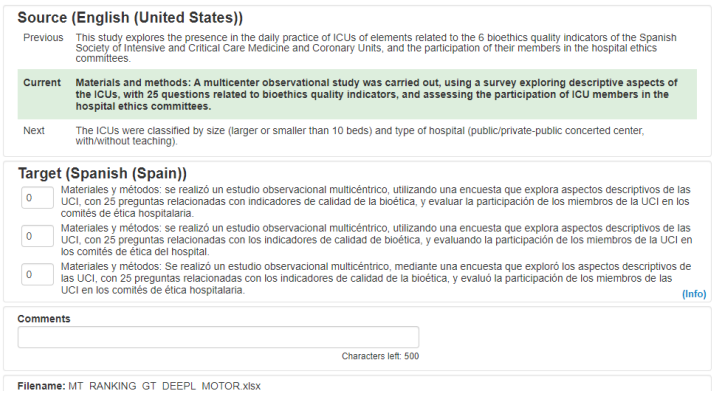
Queridos/as posteditores/as,

El objetivo principal de esta tarea es cuantificar la calidad de nuestro traductor automático. Para ello, hemos preparado una sencilla tarea de comparación de tres motores de traducción automática. A continuación, explicamos los pasos a seguir para completarla. Recordad, los textos que encontrareis son del ámbito de la **biomedicina**.

La tarea

En esta tarea nos encontramos ante 100 segmentos traducidos con tres traductores automáticos diferentes, entre ellos nuestro traductor automático basado en Marian.

- Como veréis, habéis recibido un correo electrónico de dqf@taus.net con el enlace a la prueba junto con unas breves instrucciones, que no difieren de las que encontramos aquí.
- Una vez hecho clic en el enlace, llegaréis a una página en la que podéis hacer clic en "**Start**" para empezar la evaluación.
- A continuación, encontraréis en pantalla el primer segmento, cuya estructura será la misma para todos los que lo siguen. Encontrareis el "**source**" con la oración original en inglés i **tres "targets**" con las traducciones de los tres motores. No sabréis a qué motor corresponde cada traducción, lo cual hace que el análisis sea completamente objetivo (aparte de que en cada segmento de distribuyen de manera diferente). Veréis que la pantalla que se muestra es como la siguiente:



- Aquí tenéis que asignar a las traducciones un **número del 1 al 3**, siguiendo las instrucciones de la tabla que se muestra al final de estas instrucciones.
- Si queréis, podéis añadir un **comentario** en "comments", pero es completamente opcional.
- Es posible que algunas traducciones sean iguales o creáis que merecen la misma puntuación. Si es así, podéis asignarles la **misma puntuación**, pero intentad no abusar de ello y diferenciarlas lo máximo posible. Por ejemplo, si creemos que dos traducciones son iguales y mejores que otra, podemos asignarles dos 1 a esas traducciones y un 3 a la otra, o viceversa.
- Para pasar al siguiente segmento haced clic en "**Continue**".
- Podéis volver a segmentos anteriores haciendo clic en "**Previous**".
- Podéis interrumpir la evaluación haciendo clic en "**Home**".
- Si habéis salido del navegador y queréis volver, podéis hacer clic en el enlace del correo y este os llevará al segmento en el que os hayáis quedado.
- Una vez hayáis acabado, los resultados nos llegarán automáticamente para ser analizados.

A continuación, mostramos la tabla de clasificación:

1	Es la mejor traducción de las tres que aparecen.
2	Es la segunda mejor traducción de las tres que se ofrecen.
3	Es la peor traducción de las tres que aparecen.

¡Muchas gracias por vuestra participación y colaboración!

Marta Méndez

Annex 13

Classificació d'errors

Category	Subcategory	Critical	Major	Minor	Preferential/Improvement
Country Standards	Dates, units of measurement, currency, delimiters, addresses, phone numbers, zip codes, shortcut keys, cultural references…	0	0	1	0
Kudos	:-) Congratulations	0	0	0	0
		8	**21**	**51**	**0**
Language	Capitalization	0	2	2	0
	Grammar-Syntax	6	11	28	0
	Other	1	1	1	0
	Punctuation	1	6	19	0
	Spelling	0	1	1	0
		0	**8**	**0**	**0**
Layout	Corrupted characters	0	8	4	0
	Corrupted tags	0	0	0	0
	Formatting	0	0	0	0
	Incorrect cross-ref	0	0	0	0
	Link not working	0	0	0	0
	Missing tags/variables	0	0	0	0
	Missing/Invisible text	0	0	0	0
	String-length error	0	0	0	0
	Truncation/Overlap	0	0	0	0
		33	**31**	**27**	**1**
Meaning	Addition	2	1	0	0
	Ambiguous	0	0	1	0
	Mistranslation	20	24	15	0
	Not appropriate for context	0	0	4	0
	Omission	1	2	2	0

	Other	9	4	4	1
	Untranslated text	1	0	1	0
	Word/Tag order	0	0	0	0
Other	Other	**0**	**2**	**1**	**2**
Repeat	Repeated error	**1**	**2**	**10**	**1**
		3	**28**	**67**	**8**
	Awkward syntax	2	9	16	2
	Inconsistent with refmat	0	0	1	0
	Inconsistent within text	0	0	0	0
Style	Literal translation	0	13	20	3
	Non-compliance with SG	0	0	0	0
	Tone	0	2	13	0
	Unidiomatic use	1	4	17	3
		2	**4**	**11**	**0**
	3rd party non-compliance	0	0	0	0
	Glossary non-compliance	0	0	0	0
	Inconsistent	1	2	4	0
Terminology	Industry non-compliance	0	0	0	0
	Not appropriate for context	0	0	1	0
	Other	1	2	6	0
	Prev. correction non-compliance	0	0	0	0
	TM non-compliance	0	0	0	0
Total		**47**	**96**	**168**	**12**