Entrenament d'un motor de TAE a partir d'un corpus de guies docents

Preparació del corpus, entrenament del motor de TAE amb MTradumàtica i avaluació de qualitat

Rut Geli Anticó

Treball de fi de màster

Tutor: Ramon Piqué Huerta

Màster en Tradumàtica: Tecnologies de la Traducció

Facultat de Traducció i Interpretació, Universitat Autònoma de Barcelona, 2020

Resum

En aquest TFM es porta a terme l'entrenament d'un motor de traducció automàtica estadística del català i el castellà cap a l'anglès a partir d'un corpus de guies docents de la Universitat Autònoma de Barcelona i mitjançant la plataforma MTradumàtica. El treball consta d'una explicació teòrica, en què es parla sobre la traducció i l'avaluació automàtiques, MTradumàtica i els recursos lingüístics emprats, i d'un apartat sobre la metodologia, en què s'exposen detalladament les fases del procés: creació dels corpus, entrenament del motor i avaluació de qualitat. Es comenten i s'analitzen els resultats per determinar si les guies docents són un tipus de text adient per a l'entrenament de motors de traducció automàtica estadística.

Paraules clau: traducció automàtica, traducció automàtica estadística, guies docents, MTradumàtica.

Abstract

The aim of this project is training a statistical translation engine from Catalan and Spanish into English based on teaching guides from the university Universitat Autònoma de Barcelona, and using the platform MTradumàtica. A theoretical explanation is provided, including information about machine translation, automatic evaluation, MTradumàtica and the linguistic resources used, as well as the working methodology, which can be divided in the following processes: corpus creation, engine training and quality evaluation. The results are stated and analyzed in order to determine whether teaching guides are an appropriate text type to train statistical translation engines.

Keywords: machine translation, statistical machine translation, teaching guides, MTradumàtica.

Resumen

En este TFM se lleva a cabo el entrenamiento de un motor de traducción automática estadística del catalán y el castellano hacia el inglés a partir de un corpus de guías docentes de la Universitat Autònoma de Barcelona y mediante la plataforma MTradumàtica. El trabajo consta de una explicación teórica, en la cual se habla sobre la traducción y la avaluación automáticas, MTradumàtica y los recursos lingüísticos utilizados, y de un apartado sobre la metodología, en el cual se exponen detalladamente las fases del proceso: creación de los corpus, entrenamiento del motor y avaluación de calidad. Se comentan y se analizan los resultados para determinar si las guías docentes son un tipo de texto adecuado para el entrenamiento de motores de traducción automática estadística.

Palabras clave: traducción automática, traducción automática estadística, guías docentes, MTradumàtica.

ÍNDEX PRINCIPAL

ÍNDEX DE TAULES

ÍNDEX D'IL·LUSTRACIONS

1. Introducció

La tecnologia de la informació i la comunicació (TIC) evoluciona constantment en tots els àmbits, també en el de la traducció. En concret, la traducció automàtica (d'ara endavant, TA) és un camp en clar desenvolupament, que tot i que de moment encara està lluny de poder substituir la traducció humana, ofereix una qualitat cada vegada més alta i això fa que s'estigui implementant a molts àmbits. També es tracta d'una temàtica amb diversos objectes d'estudi d'interès, entre els quals el d'aquest treball: l'entrenament de motors de traducció.

Degut a l'augment de l'ús de TA, hi ha empreses i institucions que decideixen optar per crear el seu propi motor de traducció i entrenar-lo amb els seus textos, en funció de les seves necessitats. Com que el motor està format per textos semblants als que ha de traduir, els resultats són molt més òptims. Tal com s'afirma a l'informe ProjecTA (p.28), «a llarg termini la implantació d'un sistema adaptat amb corpus propis milloraria la productivitat». A més, també es tenen en compte altres qüestions com la confidencialitat i la protecció de dades.

La finalitat d'aquest treball és crear un motor de traducció automàtica estadística (a partir d'ara, TAE) per a la Universitat Autònoma de Barcelona (UAB). La UAB és un referent pel que fa a l'aposta pel multilingüisme i una política lingüística de qualitat, i, des del 2008, té definit un Pla de llengües: un full de ruta que assegura el desenvolupament de diverses tasques en matèria lingüística a la universitat. Dins d'aquest full de ruta, es preveu que totes les guies docents de la universitat s'ofereixin en català, castellà i anglès, i, per tant, des de la Facultat de Traducció i Interpretació s'ha considerat que una aplicació molt útil de la TA pot ser la traducció de guies docents d'assignatures.

A partir d'aquí, l'objectiu d'aquest treball és crear un corpus trilingüe de guies docents en català, castellà i anglès i entrenar un motor. Es tracta d'un projecte amb una utilitat real, fet que és una de les meves motivacions principals a l'hora de dur-lo a terme. A més, considero que és un treball que pot aportar-me molts coneixements nous i em pot servir de cara al meu futur professional, perquè una de les sortides del màster és l'assessorament a empreses en la implementació de TA.

Com a reflexió, en aquesta introducció també cal remarcar que, tot i que no és l'objectiu principal del treball, una de les finalitats és donar valor a la TA, ja que sovint els traductors, i la societat en general, som reticents a incorporar les noves tecnologies a la nostra professió. Actualment, només amb traducció manual no es podria assumir ni una mínima part del volum de textos que pot gestionar la TA, així que no només s'augmenta la productivitat, sinó que també es contribueix a la difusió del coneixement. L'ús de TA fa que la professió del traductor s'hagi de reinventar, però no en suprimeix les funcions, ja que és una figura que segueix sent necessària amb la postedició. La TA complementa la traducció humana, i viceversa.

Pel que fa a l'estructura del treball, constarà d'un marc teòric, on s'introduirà la TAE i MTradumàtica, i de l'exposició de com s'ha dut a terme la part pràctica, que consta de les fases de creació dels corpus, entrenament dels motors i avaluació de qualitat.

2. Objectius i hipòtesis

A continuació, es detallaran les hipòtesis i els objectius concrets del treball, que s'han tingut sempre presents a l'hora de portar-lo a terme i que seran fonamentals a l'hora de valorar els resultats obtinguts.

Objectius

- Crear un corpus trilingüe en català, anglès i espanyol a partir de les guies docents proporcionades per la universitat.
- Crear un motor de TAE amb MTradumàtica a partir dels corpus creats.
- Entrenar el motor i optimitzar-lo.
- Valorar els resultats finals i extreure conclusions sobre la viabilitat del motor i l'adequació dels corpus.

El treball vol fixar les directrius més adequades a seguir per determinar amb la màxima exactitud possible la configuració del motor (com s'han d'agrupar els textos, quin idioma ha de ser l'original, quin format és el més adient, etc.), perquè serveixi de referència per a altres casos similars.

Hipòtesis

- Les guies docents són un tipus de text adient per a la TA.
- Si s'utilitza la postedició de TA en lloc de la traducció humana s'augmenta la productivitat i es manté la qualitat.
- Un motor de TAE creat només amb guies docents dona més bons resultats que un TAE genèric.
- Si s'agrupen les guies docents per branques de coneixement, els resultats encara són més òptims.

3. Marc teòric

En aquest apartat, es presenta el marc teòric del treball. D'una banda, es farà una breu introducció sobre la TA i el context actual i s'aprofundirà en el funcionament de la TAE. De l'altra, es parlarà sobre MTradumàtica, la plataforma que s'ha fet servir per crear el motor.

3.1. Traducció automàtica i context actual

La TA no només tradueix paraules, sinó que trasllada estructures d'una llengua a una altra; mantenint-ne el significat, però utilitzant formes genuïnes de la llengua d'arribada. Hi ha diverses maneres de definir-la:

> «Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another» (European Association for Machine Translation).

> «La *traducció automàtica* (TA) és el procés de traducció, mitjançant un sistema informàtic (compost per ordinadors i programes), de textos informatitzats escrits enla llengua origen a textos informatitzats escrits en la llengua meta.» (Ginestí i Forcada, 2009)

El llenguatge humà és un sistema molt difícil de traslladar completament a la tecnologia, perquè no només és un conjunt d'informació lexicogramatical, sinó que també hi intervé la pragmàtica: el significat sovint s'ha d'interpretar en funció del coneixement del món, un coneixement de què no disposen els ordinadors i que fa que es produeixin ambigüitats lèxiques o estructurals. Tot i això, la TA ha evolucionat molt i actualment ha arribat a cotes de qualitat molt superiors a les esperades fa un parell de dècades.

Pel que fa als usos de la TA, es poden classificar en dos: l'assimilació i la disseminació. L'objectiu de l'assimilació és que els lectors que no tenen cap coneixement de l'idioma en què està escrit un missatge puguin captar-ne la informació general; es tracta d'una traducció purament informativa i no és publicable. La disseminació, en canvi, té una finalitat productiva: s'obté un esborrany que, un cop posteditat, esdevé una traducció amb qualitat alta que és publicable. Ho explica Sánchez-Martínez (2011):

> «Machine translation (MT) has become a viable technology that helps individuals in assimilation —to get the gist of a text written in a language the reader does not understand— and dissemination —to produce a draft translation to be post-edited for publication— tasks.»

La TA es pot aplicar de maneres diferents en funció de la qualitat desitjada pel client i la finalitat del text a traduir. Si l'objectiu és l'assimilació i només es considera important que s'entengui la idea principal del text (el que en anglès s'anomena «gist»), es pot utilitzar TA sense postedició.

Per exemple, com s'explica a l'article *Making the most of Machine translation today* (ULG, 2016), pot ser que una empresa utilitzi la TA per traduir correus electrònics i comunicats interns (comunicats de l'empresa als treballadors amb una intenció informativa) o, si té un volum molt gran de documents, per determinar quins són rellevants i quins no.

Quan l'objectiu és la disseminació, la TA va acompanyada d'una postedició. En funció del client i la finalitat del text, la postedició pot ser parcial (*light post-editing*) o total (*full post-editing*). Segons criteris de l'organització TAUS, tant un tipus de postedició com l'altra han de crear un text comprensible i precís, que mantingui el contingut principal i el significat original del missatge i n'asseguri l'objectiu comunicatiu. Així doncs, el tret que les diferencia és l'estil: un text que sigui producte d'una postedició parcial presentarà alguns errors sintàctics i gramaticals; si la postedició és total, l'estil no tindrà aquests problemes. Cal tenir present, però, que la qualitat no serà equiparable a una traducció humana.

No es recomana utilitzar TA en textos de màrqueting, per exemple, perquè contenen un llenguatge complex i que s'ha de traslladar d'una manera genuïna en la llengua d'arribada (ULG, 2016).

Actualment, tot i que la traducció manual és indispensable per camps específics en què la manca de precisió o de creativitat pot tenir conseqüències crítiques, com ara, com s'acaba de comentar, els textos de màrqueting, la TA combinada amb postedició (parcial o total) cobreix gran part de la demanda actual. A més, la productivitat que ofereix la TA fa que s'incrementi el volum de traducció assumible (Aranberri, 2014).

Precisament pel gran augment de productivitat que suposa, l'ús de la TA en les empreses de serveis lingüístics s'ha estès considerablement en els darrers anys. Sánchez-Martínez (2012) ho exposa mitjançant algunes dades:

> «El uso de sistemas de traducción automática (TA) para la producción de borradores para la posedición ha crecido en los últimos años. Así lo atestigua el informe publicado por TAUS (2009) en el que se estudiaron las prácticas en lo que respecta a la automatización del proceso de traducción de varios proveedores de servicios lingüísticos: de estos, el 40% declaró hacer uso de TA, mientras que del 60% restante, el 89% dijo tener planes para la incorporación de TA en los próximos dos años. El motivo de este incremento en la adopción de sistemas de TA es la ganancia de productividad que se obtiene, como se desprende de diversos estudios (Guerberof, 2009; de Almeida & O'Brien, 2010; Plitt &Masselot, 2010). De acuerdo con TAUS (2009) esta ganancia está en torno al 70%, con una reducción de costes de entre el 30% y el 40%.»

També cal destacar que la integració de sistemes de TA a les eines de traducció assistida és una de les causes que ha motivat l'augment de l'ús de la TA a les empreses de traducció. La TA s'integra al procés de treball i es combina amb memòries de traducció, bases de dades

terminològiques, correctors ortogràfics, etc. En una mateixa eina pots accedir a totes aquestes funcions i, a més, també hi pots dur a terme la postedició.

3.2. Tipus de TA

Hi ha dos tipus d'aproximacions a la traducció automàtica. D'una banda, la traducció automàtica basada en regles (TABR o RBMT, de *Rule-Based Machine Translation*), que es configura a partir d'un coneixement lingüístic explícit. De l'altra, la traducció automàtica basada en corpus, que es basa en corpus de textos bilingües i que inclou la traducció automàtica estadística (TAE o SMT, de *Statistical Machine Translation*) i la traducció automàtica neuronal (TAN o NMT, de *Neural Machine Translation*) (Forcada, 2009).

3.2.1. Traducció automàtica basada en regles

Des dels inicis de la TA, que va sorgir a finals dels anys cinquanta, amb els primers ordinadors, fins a principis dels anys noranta, els sistemes de TA basada en regles van suposar l'aproximació a la TA dominant. Aquest tipus de sistemes obtenen la traducció d'un text en la llengua meta a partir de diccionaris computacionals, tant monolingües com bilingües, i gramàtiques computacionals amb regles morfològiques i sintàctiques. Requereixen equips amb informàtics i experts en traducció que compilin diccionaris en forma electrònica, programin analitzadors morfològics i sintàctics i defineixin regles de transformació gramatical, entre altres accions (Forcada, 2009).

Els sistemes de TABR més habituals són els de TA per transferència, que funcionen en tres fases diferenciades: la fase d'anàlisi, en què es produeix una representació intermèdia abstracta de la frase original; la fase de transferència, que reconverteix la representació obtinguda en una nova representació intermèdia en la llengua meta, i la fase de generació, en què, a partir de la representació, es genera la frase concreta en la llengua meta (Ginestí i Forcada, 2009).

Els inconvenients principals de la TABR són el desenvolupament del sistema, que requereix temps perquè s'ha d'introduir tota la informació lingüística, i la manca de naturalitat que a vegades afecta les traduccions. Tot i això, no es necessita un programari gaire potent ni cap corpus, i els errors són molt més predictibles i, per tant, és més fàcil corregir-los:

> «a database of rules and lexical items on which the rules apply […] are «readable» and can be modified by a linguist/lexicographer». (SYSTRAN Blog, 2016)

A més, és una opció interessant per aquelles combinacions lingüístiques que no disposen de gaire corpus de traducció, com és el cas de moltes llengües minoritzades.

3.2.2. Traducció automàtica estadística

Actualment, l'aproximació a la TA dominant és la basada en corpus. Als anys noranta es van publicar els resultats dels primers experiments amb un sistema de traducció automàtica estadística i aquest tipus de sistemes ja eren els dominants a principis del segle XXI (Hutchins, 2014).

Els sistemes de TA estadística s'entrenen amb grans corpus bilingües paral·lels, a partir dels quals es configuren de manera automàtica els paràmetres de diversos models estadístics. Aquests models puntuen les possibles traduccions i s'escull la més probable. Els sistemes de TAE més utilitzats en l'actualitat són els basats en segments bilingües; en anglès, *Phrase-based Machine Translation* (PBMT). Expressat amb altres paraules:

> «Statistical MT, or PBMT, on the other hand, uses predictive algorithms to translate text. These systems are built upon parallel bilingual text corpora, which serve as a basis for "matching" to create output with the highest probability of being correct.» (ULG, 2016)

Així doncs, la TA estadística porta a terme la traducció per mitjà de models estadístics, els paràmetres dels quals es configuren de manera automàtica a partir de corpus paral·lels i monolingües (Sánchez-Martínez, 2012). El procés de treball de la TAE es basa en dos processos: l'entrenament (en anglès, *training*) i la decodificació (en anglès, *decoding*).

Tal com expliquen Hearne i Way (2011), d'una banda, el procés d'**entrenament** consisteix a extreure models estadístics de traducció a partir de corpus lingüístics. En concret, dos models específics: un model estadístic de traducció a partir d'un corpus paral·lel, que relaciona segments entre l'idioma original i l'idioma de destí, i un model estadístic de la llengua d'arribada a partir d'un corpus monolingüe (normalment més extens), que s'utilitza per avaluar la fluïdesa de les traduccions.

El model de traducció s'encarrega de calcular la probabilitat condicionada que un segment de text en la llengua meta sigui la traducció d'un segment de text en llengua d'origen. Per cada parell d'oracions alineades, calcula l'alineament a nivell de paraula i, a partir d'aquest alineament, extreu equivalents de segments bilingües. Així doncs, consisteix un diccionari bilingüe en què cada possible traducció d'un segment original concret té una probabilitat associada (Hearne i Way, 2011). Quan s'introdueix una oració nova, pot calcular la traducció més probable.

El model de llengua està constituït per grans corpus monolingües de textos en la llengua meta i proporciona la probabilitat que una cadena de text meta sigui realment una oració genuïna en la llengua en qüestió. Com exposen Hearne i Way (2011), comprèn una base de dades de seqüències de paraules (entre 1 i 7) amb la traducció corresponent i associades a una probabilitat concreta. Es tracta d'un model basat en segments formats per *n* paraules, els anomenats *n*-grames, i que

calcula la freqüència dels *n*-grames als texts d'entrenament per determinar la probabilitat de les equivalències. Gràcies a aquesta probabilitat, el sistema pot estimar la genuïnitat d'una oració.

Aquests models s'utilitzen en la **decodificació**, que és la fase en què realment es genera una traducció. Aquest procés tracta la traducció com un problema de cerca: a partir de la frase que s'ha de traduir, cerca totes les traduccions i els reordenaments que permet el model de traducció i selecciona l'opció a què s'ha assignat la probabilitat global més alta, tenint en compte els models de traducció i de llengua (Hearne i Way, 2011).

En definitiva, un sistema de TAE utilitza el model de traducció per proposar les hipòtesis de traducció més probables d'una oració concreta i assignar-hi una puntuació concreta. Després, aquestes hipòtesis es contrasten amb el model de llengua, per assegurar que la traducció sigui genuïna en la llengua d'arribada. A més, també s'hi apliquen altres models probabilístics, com ara models de traducció de segments de longitud variable, que limiten la probabilitat dels segments llargs (com que apareixen poques vegades, acostumen a tenir una probabilitat molt alta), o models de reordenament, que condicionen les posicions de les paraules (Hearne i Way, 2011).

3.2.3. Traducció automàtica neuronal

La traducció automàtica neuronal és el sistema que ha sorgit més recentment, aproximadament al 2015. Primer es van integrar models de TAN en sistemes de TAE i després ja es van desenvolupar de manera independent. Actualment, pràcticament la totalitat de la recerca en traducció automàtica està dedicada a la TAN i progressa a un ritme accelerat (Koehn, 2017).

La TAN és similar a la TAE, «but uses a completely different computational approach: *neural networks*» (Forcada, 2017). Aquest tipus de sistemes prenen com a referència les neurones del cervell humà, ja que estan formats per milers d'unitats artificials (que també anomenem «neurones») que s'activen en funció dels estímuls que reben d'altres neurones (Forcada, 2017). Els conjunts de neurones connectades formen xarxes neuronals, a partir de les quals el sistema utilitza algoritmes per aprendre automàticament patrons de regularitats en les seqüències de paraules i agrupar valors per formar representacions d'informació (ULG, 2016). Aquest procés d'aprenentatge automàtic es coneix com a *deep learning*.

En general, el procés de treball de la TAN es divideix en dues fases: la codificació, en què el sistema analitza les representacions de les paraules de la frase original per produir una representació vectorial, i la descodificació, en què, a partir d'aquesta representació, el sistema prediu les paraules en la llengua meta (Casacuberta, Peris; 2017).

Els sistemes de TAN es basen en grans quantitats de corpus bilingües i requereixen tant un maquinari molt potent com un període de temps elevat per a l'entrenament. Ho exposa Forcada (2017):

> «NMT usually requires very large training corpora, typically as large as those used in good old SMT, and its training (searching for the best value for all of the weights in the network) is computationally very demanding: most NMT training resorts to using dedicated number-crunching hardware evolved from graphics processors, with typical training times ranging from days to months.»

A més, la TAN sovint genera errors semàntics; un tipus d'error que és particularment greu, perquè ocasiona canvis en les paraules amb un rol clau pel que fa al significat global de la frase. Per exemple, quan troba paraules que no coneix, fa traduccions parcials o utilitza paraules similars, com ara escriure un país en lloc d'un altre (Forcada, 2017).

Si comparem els sistemes de TAN amb els de TAE:

> «NMT employs artificial intelligence algorithms that can derive meaning from whole sentences or ideas using so-called neural networks whereas PBMT works with individual words, or segments of a sentence.» (Wu, et all., 2016)

3.2.4. Elecció del motor de TA

Tenint en compte els avantatges i els inconvenients de cada tipus de sistema, s'ha determinat que el més adequat als objectius d'aquest treball és el de TAE. S'ha arribat a aquesta conclusió en base a diversos motius, que s'exposen a continuació.

S'ha descartat l'aproximació a la TA basada en regles perquè implica codificar a mà tota la informació lingüística, que, com ja s'ha comentat, es tracta d'un procés molt costós i que requereix molt de temps. Tot i que es disposen de coneixements lingüístics pel que fa als idiomes involucrats al sistema, no es disposa del temps suficient per introduir les regles sintàctiques i les dades lèxiques per definir el motor.

> «Statistical methods do not require researchers to know the languages involved in systems (or, at least, to have indepth knowledge) and do not demand complex large-scale acquisition of rules and lexical data.» (Hutchins, 2014)

També intervé en la decisió el tipus de text que es vol traduir. Com ja s'ha anat comentant i tal com s'amplia més endavant, l'objectiu del treball és crear un motor que pugui traduir guies docents d'assignatures universitàries: un camp molt específic i un tipus de text amb una estructura i unes característiques molt concretes. És en casos com aquest, en què la finalitat és traduir textos tan específics i, a més, es disposa d'un corpus molt extens (de més de 350.000 línies), quan l'aproximació a la TA basada en corpus funciona més bé:

> «In corpus-based approaches of machine translation, the more specific the training corpus domain, the better the translation output will be.» (Doğru, Martín-Mor i Aguilar-Amat, 2018)

Si tenim això en compte, es pot tirar entre crear un motor de TA estadística o bé neuronal. Tot i això, la TAN s'ha descartat, perquè no es disposa del maquinari adient. Com ja s'ha comentat a l'apartat anterior, un requisit indispensable per entrenar un motor d'aquest tipus és tenir quantitats ingents de dades i processadors gràfics molt potents.

En definitiva, la disponibilitat de temps i de recursos i l'especificitat de l'àmbit de què es volia crear el motor han suposat els criteris principals que s'han tingut en compte a l'hora d'escollir quin tipus de motor s'entrenava.

3.3. Avaluació automàtica de la qualitat de la TA

L'avaluació automàtica de la qualitat de la TA és una fase indispensable en l'entrenament d'un motor de TAE, principalment per dos motius: poder analitzar la qualitat dels resultats del motor i poder-ne ajustar els paràmetres. Pel que fa a l'ajust dels paràmetres, a partir dels resultats que proporcionen les mètriques d'avaluació automàtica de la qualitat de la TA, es pot reassignar el pes de cadascun dels models probabilístics (models de traducció de segments de longitud variable, models de reordenament, etc.) i, d'aquesta manera, optimitzar automàticament el sistema (Koehn, 2010).

Tal com exposen Papineni et al. (2002), l'avaluació humana de la TA és molt més precisa i ofereix avaluacions més completes, però també és cara; és per això que s'han dissenyat sistemes d'avaluació automàtica. Per determinar la qualitat d'una traducció automàtica, aquests sistemes d'avaluació automàtica necessiten una mètrica numèrica de «translation closeness» i un corpus de traduccions humanes de referència de qualitat.

Els tres mètodes d'avaluació automàtica de la TA més emprats actualment són: BLEU, METEOR i TER. A continuació, s'exposen breument, tot i que per a la realització d'aquest treball només s'ha emprat el BLEU.

3.3.1. BLEU

El mètode BLEU (Papineni et al., 2002), que prové de les sigles en anglès *Bilingual Evaluation Understudy*, mesura la distància entre una traducció feta amb TA i una de referència mitjançant *n*-grames: seqüencies d'entre 1 i 4 paraules. A partir de la traducció humana de referència, mesura el nombre de coincidències amb la traducció feta amb TA (Babych, 2014).

Tal com exposen Papineni et al. (2002), a partir d'aquestes coincidències, s'assigna a la TA un nombre del 0 a l'1, essent 0 la puntuació si no coincideix cap paraula i 1 si coincideixen totes. Tret que la TA sigui idèntica a la traducció de referència, és molt complicat que el resultat sigui 1:

> «The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. For this reason, even a human translator will not necessarily score 1.» (Papineni et al., 2002)

Això es deu al fet que no es tenen en compte les coincidències aproximades, com ara els sinònims.

3.3.2. METEOR

El mètode METEOR (Banerjee and Lavie, 2005) es tracta d'una modificació de la mètrica BLEU, ja que també funciona mitjançant *n*-grames, però se'n diferencia perquè integra funcions lingüístiques addicionals, com ara sinònims o estructures equivalents, tal com exposa Babych (2014):

> «[…] a metric which integrates additional linguistic features, such synonyms and stems, or dictionary forms of the inflected words found in the evaluated texts.»

Segons Lavie i Denkowski (2009), la mètrica METEOR, a diferència de BLEU, té en compte les variants morfològiques i els sinònims com a correspondències vàlides. A més, no només es basa en la precisió (la proporció de paraules de la TA que coincideixen amb la traducció de referència), sinó que també inclou la cobertura o *recall* (la proporció de paraules de la traducció de referència que coincideixen amb la TA), una propietat que té una correlació molt alta amb els criteris humans:

> «In contrast with IBM's BLEU, which uses only precision-based features, METEOR uses and emphasizes recall in addition to precision, a property that has been confirmed by several metrics as being critical for high correlation with human judgments.» (Lavie i Denkowski, 2009)

3.3.3. TER

La mètrica TER (Snover et al. 2006), en anglès *Translation Error Rate*, mesura la quantitat de canvis que ha de fer un posteditor humà per aconseguir que la TA correspongui amb una traducció de referència:

> «TER measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation» (Snover et al. 2006)

És a dir, segons Babych (2014), aquest mètode té en compte variacions de posició, i obtén el resultat calculant el nombre d'edicions que es farien i dividint-lo entre el nombre mitjà de paraules d'una de les traduccions de referència. 0 és el valor òptim.

3.4. Recursos lingüístics

Els recursos lingüístics que s'han utilitzat per crear el corpus són textos proporcionats per la Universitat Autònoma de Barcelona. Com ja s'ha comentat, es tracten de guies docents d'assignatures universitàries i es van obtenir a través de la Facultat de Traducció i Interpretació; en concret, Pilar Sánchez Gijón i Ramon Piqué Huerta es va encarregar de fer les gestions corresponents per demanar els textos.

Es van rebre tres tipus de recursos: les guies docents, les competències i els resultats d'aprenentatge.

3.4.1. Guies docents

Les guies docents conformaven la major part dels arxius i constaven d'unes 350 000 línies. Es van proporcionar ordenades per carpetes codificades i cadascuna contenia tres fitxers: les versions en català, en espanyol i en anglès. Cada carpeta es tractava d'una assignatura, però no hi constava de quina assignatura concreta es tractava i els codis no seguien cap criteri definit. Els fitxers estaven en format HTML i codificats amb ANSI.

Un cop s'obria l'arxiu, gairebé totes les guies seguien l'estructura que s'indica a continuació, tot i que algunes suprimien algun apartat (per exemple, el de prerequisits):

- Títol de l'assignatura: s'indica el codi i els crèdits de l'assignatura, juntament amb el nom, el tipus, el curs i el semestre de la titulació.
- Professor/a de contacte: s'indica el nom i el correu electrònic del professor.
- Utilització d'idiomes a l'assignatura: es defineix la llengua vehicular i si hi ha grups íntegres en altres idiomes.
- Equip docent: s'indiquen els professors.
- Prerequisits: s'estableixen uns requisits que se sobreentén que l'estudiant que s'inscriu a l'assignatura compleix.
- Objectius: s'inclouen els objectius principals de l'assignatura.
- Competències: les competències que necessitarà l'estudiant.
- Resultats d'aprenentatge: què s'espera que els estudiants aprenguin.
- Continguts: s'indica com s'estructuren els continguts de l'assignatura.

- Metodologia: es detalla com funciona l'assignatura: quin tipus de classes i d'avaluació es porta a terme, les hores de dedicació de l'estudiant, etc.
- Activitats formatives: es complementa l'apartat anterior amb una informació més esquemàtica sobre el tipus de classes que es porten a terme, els crèdits que compten i quins resultats d'aprenentatge s'assoliran, entre d'altres.
- Avaluació: s'explica detalladament com s'avaluarà l'assignatura.
- Activitats d'avaluació: informació esquemàtica relacionada amb l'apartat anterior, on es concreta el percentatge que val cada activitat avaluable, els crèdits que compta i els resultats d'aprenentatge que s'assoleixen, entre d'altres.
- Bibliografia: s'indica la bibliografia de l'assignatura.

Pel que fa a la tipologia textual i tal com s'acaba de comentar, les guies docents segueixen una estructura molt marcada i sovint hi ha estructures sintàctiques que es repeteixen. Els textos amb estructures senzilles i de tipus més aviat tècnic són els més idonis per a la TA, tal com demostren diversos estudis:

> «The findings show that machines produce better translations of technical sets of instructions then of other types of texts.» (Calude, Andreea, 2003)

3.4.2. Competències i resultats d'aprenentatge

Dins de cada guia docent, tal com s'acaba de comentar, hi ha un apartat sobre les competències que necessita l'estudiant per cursar l'assignatura i un altre sobre els resultats d'aprenentatge que s'espera que assoleixi. Tant les competències com els resultats estan formats per frases preestablertes que s'utilitzen en tota la universitat.

Aquests conceptes estan traduïts per professionals perquè els docents puguin incorporar tant la versió en català com en castellà o anglès quan facin la versió en cada idioma de la guia docent. Així doncs, com que les equivalències són de qualitat, tot i que conformen una part molt petita del corpus, en fan augmentar el nivell de qualitat.

A més, també hi ha oracions senceres repetides, sobretot en l'apartat sobre les competències, ja que s'hi descriuen unes competències que són universals per a tota la universitat. Aquestes competències ja estan traduïdes i els docents poden consultar-les en tots tres idiomes en un full de càlcul. Tot i això, no es té constància de cap sistema automatitzat que les insereixi a les guies docents, sinó que es fa de manera manual.

3.5. MTradumàtica

Una de les eines més populars per a la construcció i l'ús de sistemes de TAE (Koehn, 2009) és Moses: una plataforma de codi obert, però que aconsegueix resultats comparables en qualitat i eficiència als sistemes privatiu. Amb Moses, els usuaris no s'han de preocupar de programar un sistema de TAE propi. Malgrat tot, per administrar Moses cal tenir coneixements de sistemes UNIX i del terminal, perquè no disposa d'una interfície gràfica d'usuari, fet que suposa una barrera per a molts usuaris (Martín-Mor, Piqué, 2017). És per això que han sorgit diversos programes més accessibles que pretenen ser un complement per a Moses.

Entre aquests programes, cal destacar MTradumàtica, la plataforma de programari lliure que s'ha utilitzat per fer aquest projecte i que ha desenvolupat el grup Tradumàtica de la UAB. Es tracta d'una distribució de Moses que té com a objectiu l'entrenament i l'ús de motors de TAE. Va dirigida a investigadors i a estudiants no experts en l'entrenament i l'ús de motors de TAE que vulguin adquirir coneixements sobre el tema, i també a empreses que vulguin donar-hi una utilitat real i crear el seu propi motor de traducció:

> «L'objectiu de l'eina és proporcionar a investigadors no experts en tecnologia una aplicació web que permeti crear un motor de traducció amb Moses. L'aplicació vol servir, també, com a prova de concepte per a les empreses de traducció i els posteditors que vulguin posar a prova un flux de treballa amb motors propis de TA, sense oblidar el vessant didàctic en el marc de la docència de processos de TA adreçada a estudiants de traducció.» (Martín-Mor, Piqué, 2017)

MTradumàtica consisteix en una interfície web on es poden penjar corpus, portar a terme l'entrenament de sistemes i, a més, posar-los en funcionament fent traduccions dins mateix de la plataforma. Tal com expliquen Martín-Mor i Piqué (2017), MTradumàtica es va crear amb tres objectius: primer de tot, desenvolupar una interfície gràfica per a la plataforma per facilitar-ne l'ús als usuaris no experts i també per utilitzar l'aplicació per a finalitats educatives; el segon objectiu era que fos multiplataforma i que s'hi pogués accedir via web per tal d'evitar instal·lar res en local, i el tercer fer una instal·lació en els servidors propis de la universitat per motius de confidencialitat.

Respecte a Moses en versió terminal, MTradumàtica ofereix diversos avantatges. En primer lloc, i com ja s'ha comentat, disposa d'una interfície gràfica, que es combina amb diverses funcions addicionals que faciliten l'entrenament i l'ús dels sistemes de TAE. A més, es pot utilitzar en qualsevol sistema operatiu, perquè es proporciona una màquina virtual de VirtualBox, i també permet integrar l'aplicació en altres aplicacions. Per últim, i tenint en compte la finalitat educativa de l'eina, es poden observar fàcilment tots els passos d'un procés de traducció.

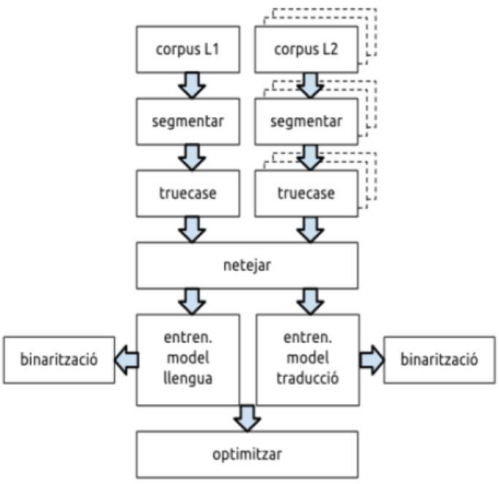
3.5.1. Procés d'entrenament amb MTradumàtica

Tal com expliquen Martín-Mor i Piqué (2017), per iniciar el procés de treball amb MTradumàtica, es necessita un corpus paral·lel bilingüe, que conformarà el motor de traducció, i un o més corpus monolingües, per formar el model de llengua.

Els primers processos que es porten a terme (automàticament) són la segmentació, el *truecasing* i la neteja dels corpus. El procés de segmentació consisteix a separar amb espais les paraules dels signes de puntuació, ja que «aïllar la puntuació permet incrementar les probabilitats d'obtenir coincidències amb els futurs textos que es traduiran automàticament» (Martín-Mor i Piqué, 2017). La fase de *truecasing* determina com és més probable que comenci una paraula: si en majúscula o en minúscula. En funció d'això, hi manté la majúscula o la minúscula; els noms propis, per exemple, es queden en majúscula. Pel que fa a la neteja dels corpus, es porta a terme «la supressió de les frases llargues i mal alineades dels corpus amb l'objectiu de minimitzar els problemes en la fase d'entrenament» (Martín-Mor i Piqué, 2017).

A continuació, el sistema processa la informació lingüística que s'ha proporcionat en la fase d'entrenament, en què, «a partir de l'anàlisi de coocurrències de paraules i segments en les dues llengües, s'infereixen de manera automàtica correspondències de traducció». L'últim procés que es porta a terme és l'optimització o *tuning*, que determina automàticament els valors òptims d'una sèrie de paràmetres perquè el motor generi les millors traduccions possibles (Martín-Mor i Piqué, 2017).

Martín-Mor i Piqué (2017) il·lustren aquest procés d'entrenament mitjançant aquest esquema:



Il·lustració 1. Esquema de procés d'entrenament amb MTradumàtica (Martín-Mor i Piqué, 2017).

3.5.2. Interfície

Els set passos dels menús de la pàgina inicial de la plataforma volen reflectir el procés d'entrenament d'un motor de TAE:

1. Carregar fitxer (*Upload files*)
2. Crear monotextos (*Create Monotexts*)
3. Entrenar models de llengua (*Build Translators*)
4. Crear bitextos (*Create Bitexts*)
5. Entrenar traductors automàtics (*Build Translators*)
6. Traduir (*Translate*)
7. Avaluar (*Inspect*)

Primer de tot, s'han de carregar al sistema els fitxers a partir dels quals es generaran els models de llengua i el de traducció. S'admeten els tipus de fitxers següents: TXT monolingües, TXT bilingües alineats per a Moses i TMX. Els arxius alineats se separen per llengües automàticament i es generen dos arxius diferents. A continuació, es crea un monotext buit i s'hi afegeixen tots els fitxers monolingües que es vulguin, sempre tenint en compte que només s'han de crear monotextos en la llengua d'arribada del motor, perquè serviran per generar els models de llengua. Al pas següent, es creen un o més models de llengua a partir del monotext que hem creat. Tot seguit, es crea el bitext seguint un procés similar al del monotext: primer es crea el bitext buit i s'hi van afegint els fitxers que es vulguin dels que hem penjat. En aquest punt ja es pot generar el motor de traducció a partir del bitext i del model de llengua. Un cop finalitzat l'entrenament, es pot aplicar-hi un procés d'optimització, que té com a objectiu assolir la màxima qualitat possible, i que s'ha de fer a partir de bitextos de la mateixa combinació lingüística, però que no formin part del corpus d'entrenament. Els processos d'entrenament del motor i d'optimització poden ser molt lents. A continuació, a la pestanya Translate, es poden traduir frases, documents o TMX amb el motor que s'ha creat i també vincular-lo amb OmegaT a partir d'un URL (Martín-Mor i Piqué, 2017).

MTradumàtica també ofereix l'opció d'avaluar motors de TA a la pestanya Avaluar; una funció en desenvolupament que permet calcular mètriques automàtiques de rendiment de la TA. Hi ha previst implementar dues possibilitats d'avaluació: comparar MTradumàtica i motors de TA externs a partir d'una traducció humana de referència, o bé directament avaluar MTradumàtica també amb un exemple de traducció humana.

4. Metodologia

A continuació, s'exposen les diverses fases de la part pràctica del treball. Bàsicament, el procés es pot dividir en la creació del corpus, l'entrenament del motor i l'avaluació de qualitat; tot i això, la part del treball en què s'ha dedicat més temps ha estat en la creació del corpus, ja que s'han hagut de netejar i reorganitzar diverses vegades els recursos lingüístics originals. Dins de cada apartat, es presenten les dificultats que ha comportat i quines solucions s'han implementat.

4.1. Recursos lingüístics i fase de proves

Els recursos lingüístics que es volien utilitzar per al corpus estaven clars des de bon començament, ja que, com s'ha comentat, la finalitat del treball era entrenar un motor amb guies docents de la UAB. Així doncs, es van fer tots els tràmits per obtenir-les.

Se'ns van enviar tres corpus:

- Corpus de guies docents
- Corpus de competències
- Corpus de resultats

Es va plantejar la possibilitat d'afegir altres recursos lingüístics, principalment en anglès, perquè el corpus fos més extens i, per tant, donés més bons resultats. Tot i això, com que es volia obtenir un motor de traducció estadística especialitzat, l'objectiu era crear-lo a partir d'un corpus format íntegrament per guies docents i, si es volia complementar amb altres recursos, calia que també fossin guies docents. Es van fer cerques a llocs web d'universitats de parla anglesa per veure si es podia baixar fàcilment un volum considerable de guies docents d'assignatures, però era una tasca difícil i es va decidir deixar-ho per a més endavant. Primer s'avaluarien els resultats del motor format exclusivament per guies docents de la UAB i es deixaria oberta l'opció d'ampliar el corpus contactant amb alguna universitat de parla anglesa per obtenir els recursos directament.

Així doncs, els primers passos van ser analitzar els arxius. Abans de rebre totes les guies disponibles, ens van enviar diverses proves per veure si el format dels arxius era l'adequat i també per determinar si realment els recursos servien per entrenar el motor. Amb aquestes proves es van fer les primeres alineacions, que van servir per prendre algunes decisions.

Una d'aquestes decisions era determinar quin programari s'utilitzaria per fer l'alineació definitiva. Com que disposàvem de moltes guies i, a més, l'extensió del corpus és un dels aspectes que influeix més en els resultats a l'hora d'entrenar un motor (com més extens és un corpus, més alta és la qualitat de les traduccions que s'obtenen), era prioritari cercar un programa que permetés

alinear directoris sencers en lloc d'arxius separats (funció *batch align*). Es va cercar informació sobre programari que oferís aquesta funció i, finalment, es va optar per LF Aligner.

En principi, la interfície del programa sembla que no permet automatitzar les alineacions de diversos arxius, però a les instruccions s'explica com crear un arxiu executable (.bat) a partir d'un Excel. Cal definir tots els paràmetres que necessita LF Aligner, tal com es mostra a continuació:

LF_aligner_4.21.exe --filetype="h" --infiles="[ruta_arxiuCA]","[ruta_arxiuEN]","[ruta_arxiuES]" --languages="ca","en","es" --segment="y" --review="xn" --tmx="n"

Es van portar a terme proves amb diversos arxius, per les quals es van haver d'ordenar els noms dels arxius i els de les carpetes per poder automatitzar la creació dels paràmetres que s'han indicat de manera senzilla des d'Excel. Es va crear un document a Notepad++ amb totes les línies que indicaven cada conjunt d'arxius i, a continuació, es va executar.

Un cop fetes les alineacions, van sorgir dos problemes. El primer estava relacionat amb l'accentuació, ja que ni en català ni en castellà es mostraven els accents a la interfície de LF Aligner ni als arxius que s'obtenien, com es pot observar a les imatges següents:

3	Advising and managing in terms of social security, social welfare and complementary social protection.	Assessorar i gestionar en mat\xE8ria de seguretat social, assist\xE8ncia social i protecci\xF3 social complement\xE0ria.
4	Distinguishing the special needs of labour integration in different groups of workers (with mental or psychical disabilities, immigrants...).	Distingir les necessitats especials d'inserci\xF3 laboral de diferents grups de treballadors (amb discapacitats f\xEDsiques o ps\xEDquiques, immigrants, etc.).

Il·lustració 2. Problemes d'accentuació a la interfície de LF Aligner.

```
2019/2020   2019/2020   2019/2020   public_ca
Gen\xF2mica Genomics    Gen\xF3mica public_ca
Codi: 42399 Cr\xE8dits: 12  Code: 42399 ECTS Credits: 12    C\xF3digo: 42399 Cr\xE9ditos ECTS: 12   public_ca
Titulaci\xF3 Tipus Curs Semestre    Degree Type Year Semester   Titulaci\xF3n Tipo Curso Semestre   public_ca
4313473 Bioinform\xE0tica / Bioinformatics  4313473 Bioinformatics  4313473 Bioinform\xE1tica / Bioinformatics  public_ca
OT  OT  OT  public_ca
0   0   0   public_ca
1   1   1   public_ca
```

Il·lustració 3. Problemes amb l'accentuació als fitxers obtinguts.

A l'arxiu *readme* de LF Aligner s'exposa que a Windows el programa no mostra els accents a la interfície, però s'afirma que als arxius que es creen sí que es mostren. Es va intentar canviar el format dels arxius i passar la codificació a UTF-8, però el programa continuava donant aquest problema als arxius que s'obtenien alineats.

El segon problema estava relacionat amb la qualitat de les equivalències. Com que les guies docents no eren textos pensats per ser alineats, sovint les equivalències no eren gaire bones, ja que algunes guies docents presentaven variacions entre les diverses versions. Les equivalències errònies entre segments eren degudes principalment al següent:

- Frases desordenades: a les guies docents, és molt freqüent trobar la informació organitzada en llistats. Pel motiu que sigui, a l'hora de fer les traduccions, les afirmacions de diversos llistats es van desordenar i, per tant, això comportava alineacions incorrectes.
- Versions resumides: sobretot en les versions en anglès, es va observar que la informació s'havia resumit en multitud de casos.
- Desplaçaments: també es va constatar que a vegades LF Aligner detectava com a truncacions elements que no ho eren (per exemple, una ela geminada) i això feia que es produís un desplaçament en una versió concreta i que la línia en qüestió ja no coincidís amb les altres.

Així doncs, en aquesta primera fase ja es va observar que la qualitat dels recursos lingüístics no era gaire bona, sobretot en la versió en anglès.

4.2. Creació dels corpus

4.2.1. Primeres alineacions i estudi comparatiu

Tenint en compte els problemes esmentats a l'apartat anterior, es va decidir demanar la col·laboració del professor de la UOC Antoni Oliver González, que ja ha participat en alguns projectes del Grup Tradumàtica. Per mitjà del llenguatge de programació Python, Oliver va alinear totes les guies automàticament i també va portar a terme una primera neteja.

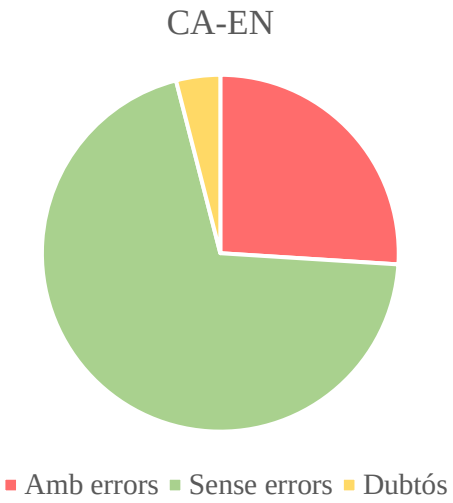
A partir del corpus obtingut, es va decidir portar a terme un estudi comparatiu per determinar d'una manera més exacta quina proporció de les equivalències tenien algun tipus d'error. Per fer-ho, es van seleccionar aleatòriament 100 segments, 50 de la memòria de traducció del català a l'anglès i 50 més de la memòria de l'espanyol a l'anglès. En un full de càlcul, es van avaluar els segments un per un tot marcant si contenien algun tipus d'error o no. Algunes equivalències es van valorar com a dubtoses, ja que contenien algun tipus d'error, però es tractava d'una qüestió sense massa importància.

A continuació, es presenten els resultats de les dues comparatives, que s'inclouen en la seva totalitat als annexos del treball.

◊ Arxiu als annexos: Estudi comparatiu > Mostres_errors.xlsx

Pel que fa a la combinació del català a l'anglès:

CA-EN		
Amb errors	13	26%
Sense errors	35	70%
Dubtós	2	4%

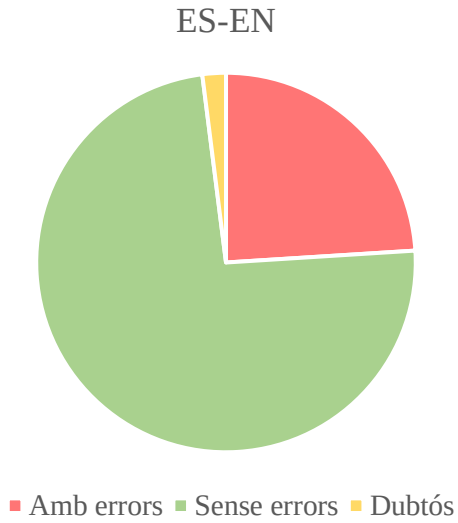


Il·lustració 4. Estudi comparatiu d'equivalències (CA-EN).

Es pot observar que un la major part de les equivalències no contenia cap error. Tot i això, un 30% eren incorrectes.

Els resultats de la comparativa de l'espanyol a l'anglès van ser similars:

ES-EN		
Amb errors	12	24%
Sense errors	37	74%
Dubtós	1	2%



Il·lustració 5. Estudi comparatiu d'equivalències (ES-EN).

El percentatge de segments sense errors va augmentar en la comparativa de l'espanyol a l'anglès. Malgrat això, si tenim en compte que es tracta d'una mostra i que, per tant, són valors aproximats, podem concloure que els resultats són molt semblants a la comparativa amb el català.

4.2.2. Processos automatitzats de neteja

Així doncs, després de constatar que aproximadament una quarta part de les equivalències del corpus contenien algun tipus d'error, es va decidir tornar a netejar els corpus amb un programa de Python, aquesta vegada amb l'ajuda de Ramon Piqué.

Es va utilitzar el programa de Python **guietoalign.py** desenvolupat per Antoni Oliver, que converteix tots els arxius HTML a text i els segmenta. A més, crea l'arxiu align.sh, que alinea totes les guies. L'script del programa s'inclou als annexos.

◊ Arxiu als annexos: guiestoalign.py

Més endavant també es van utilitzar comandes de Linux per eliminar entrades duplicades i concatenar els arxius; és a dir, convertir diversos arxius en un de sol.

La sintaxi de Linux per concatenar arxius és la següent:

cat arxius.* > arxiunou.txt

I aquesta és la sintaxi per eliminar entrades duplicades:

cat arxiunou.txt | sort | uniq > arxiunou-unic.txt

Després d'aquests processos, s'aconseguien arxius TXT com ara el següent:

```
10.- Poultry farming.   10.- Avicultura.
10.- Production of transgenic plants and applications.  Tema 10.- Producción de plantas transgénicas y aplicaciones.
10. ~~~ Serological diagnosis of infectious diseases    Diagnóstico serológico de las enfermedades infecciosas.
10 short-answer questions (10x2 points per question = 20p)  Preguntas curtas 10 (10 x 2 puntos por pregunta = 20 p)
10th Edition.   10th Edition.
10th Ed. Saunders   10th Ed. Saunders
1.10 Gas exchange.  1.10- Intercambio gaseoso.
11.1 Basis of the cell cycle control: intracellular signals and extracellular signals, control points along the cell cycle, cycles
11.1 Concept of civil and professional liability, 11.2. 11.1 Concepto de responsabilidad civil y profesional, 11.2.
11.1 Definition, aetiology and classification.  11.1. ~~~ Definición, etiología y clasificación.
1.11 Regulation of breathing.   1.11- Regulación de la respiración.
11.2 Control point at the end of G2 (G2 / M).   11.2 Punto de control al final de G2 (G2 / M).
1.1.2 Eukaryotic chromosome replication machinery.  1.1.2 Maquinaria de replicación de los cromosomas eucariotas.
1/12 group: participated in a clinical session  Clases prácticas de seminario en grupo pequeño
11.3 Exploration of normal and pathological baby.   11.3. ~~~ Exploración del bebé normal y del bebé patológico.
11.3 Output control of M    11.3 Control de salida de M
11.4 Control of the cycle at the end of G1  11.4 Control del ciclo al final de G1
11.5 Molecular brakes of the cycle  11.5 Frenos moleculares del ciclo
11.5 Post-surgical physiotherapy treatment. 11.5. ~~~ Tratamiento de fisioterapia posquirúrgico.
11.5 Testimony of the legal expert. 11.5 El testimonio de los peritos.
11.6.1 Cancer-related genes.    11.6.1 Nada relacionados con cáncer.
11.6.2 Viruses and cancer.  11.6.2 Virus y cáncer.
11.6.3 Cancer diagnosis and treatment   11.6.3 Diagnóstico y tratamiento del cáncer
11.6 Disobeying the social control of cell proliferation.   11.6 Desobedecer el control social de la proliferación celular.
11ª edición     11ª edición
```

Il·lustració 6. Exemple d'arxiu TXT després dels processos automatitzats de neteja.

4.2.3. Redefinició dels motors

També va ser en aquest estadi del procés, un cop ja s'havia treballat a fons amb els corpus en qüestió, quan es va poder concretar com s'abordaria l'entrenament del motor i en quins corpus se centraria l'estudi.

En primer lloc, es va decidir agrupar les guies docents per branques del coneixement i crear motors de traducció diferents, en base a la hipòtesi que s'obtindrien més bons resultats traduint les guies amb un motor entrenat amb un corpus format únicament per recursos lingüístics de l'àmbit en qüestió. Es van classificar les titulacions de la UAB, tal com es mostra als annexos, i es van definir les 6 branques del coneixement següents:

1. Arts i humanitats
2. Biociències
3. Ciències
4. Ciències de la salut
5. Ciències socials i jurídiques
6. Enginyeries

◊ Arxiu als annexos: Estudi comparatiu > Branques del coneixement per titulacions.pdf

Així doncs, calia agrupar les guies docents de cada branca del coneixement. Tal com s'ha explicat a l'apartat Recursos lingüístics, els recursos inicials tenien com a nom un codi que no coincidia amb el codi de l'assignatura; per tant, calia cercar cada equivalència. Es disposava d'un full de càlcul de referència, adjuntat als annexos, en què constava a quina assignatura i a quin grau pertanyia cada codi de guia docent. L'objectiu era cercar si els codis dels recursos estaven ordenats segons algun criteri concret, perquè, en cas contrari, la separació de les guies per branques no era factible, ja que separar els arxius un per un de manera manual hagués suposat una tasca massa laboriosa. Efectivament, les guies, majoritàriament, estaven ordenades per titulacions i, per tant, es va anar comprovant a quina branca del coneixement pertanyia cada estudi i es van anar distribuint els 11 477 TXT en carpetes.

◊ Arxiu als annexos: Estudi comparatiu > Codis titulacions.xlsx

Quan aquest procés va estar enllestit, es van separar cinc guies docents de cada tipus que no s'inclourien en el corpus d'entrenament i que més endavant servirien per fer l'avaluació de qualitat. A continuació, es van utilitzar les comandes de Linux detallades a l'apartat anterior per unir totes les altres guies que havien de formar el corpus d'entrenament i es van crear TXT genèrics per branques del coneixement i per idiomes. Per tant, es van obtenir els següents arxius:

corpus-unic-biociencies-en-ca
corpus-unic-biociencies-en-es
corpus-unic-biociencies-es-ca
corpus-unic-ciencies-en-ca
corpus-unic-ciencies-en-es
corpus-unic-ciencies-es-ca
corpus-unic-enginyeries-en-ca
corpus-unic-enginyeries-en-es
corpus-unic-enginyeries-es-ca
corpus-unic-humanitats-en-ca
corpus-unic-humanitats-en-es
corpus-unic-humanitats-es-ca

corpus-unic-salut-en-ca	corpus-unic-socials-en-ca
corpus-unic-salut-en-es	corpus-unic-socials-en-es
corpus-unic-salut-es-ca	corpus-unic-socials-es-ca

◊ Arxius als annexos: TXT de la carpeta Corpus per àmbits, recomptes i resultats

4.2.4. Recompte de segments i neteja de bibliografia

Un cop obtinguts els TXT unificats de cada branca del coneixement, es va decidir crear un full de càlcul amb tots els segments per analitzar-ne diversos aspectes. Abans de crear-lo, però, es van suprimir les línies inicials i finals d'alguns TXT perquè contenien només guions i frases en xinès i en àrab.

Tot seguit, es va crear un Excel de "recompte" de cada arxiu, que contenia una taula principal amb les dades següents:

- Segment original
- Traducció
- Nombre de paraules de l'original
- Nombre de paraules de la traducció

Com que l'objectiu era identificar els segments amb una qualitat alta, es va considerar que un bon criteri era calcular la similitud entre l'original i la traducció i descartar les equivalències en què hi hagués una diferència notable entre les dues frases, ja que probablement es tractaven de traduccions amb més o bé amb menys informació del compte. Es va fer una distinció entre els segments de 10 paraules o més i els de menys de 10 paraules, perquè en els més llargs es va calcular el percentatge, mentre que en els més curts es va valorar la diferència concreta de paraules (el percentatge s'alterava molt més fàcilment en aquests casos).

Així doncs, també es van incloure aquests valors a la taula:

- Percentatge de similitud en nombre de paraules
- Ponderació per frases menors de 10 paraules

Al lateral, es van incloure valor resumits dels següents aspectes:

- Total de segments de 10 o més paraules
  - ➔ Per sota de 90% de similitud en nombre de paraules
  - ➔ Entre 90% i 110% de similitud en nombre de paraules
  - ➔ Per damunt de 110% de similitud en nombre de paraules
- Total de segments de menys de 10 paraules

- ➔ Per sota d'una paraula de diferència (-1)
- ➔ Entre -1 i +1 paraula de diferència
- ➔ Per damunt d'una paraula de diferència (+1)

- Distribució dels segments per paraules i percentatge que representen sobre el total

Es van considerar segments de bona qualitats els que tenien entre un 90 i un 110% de similitud, en el cas dels segments de 10 o més paraules, i els que tenien entre -1 i +1 paraula de diferència, en el cas dels de menys de 10 paraules.

◊ Arxiu als annexos: Corpus per àmbits, recomptes i resultats > corpus-unic-[salut-en-ca]_recompte.xlsx

* Cal triar l'opció que convingui en cada cas per als noms entre claudàtors.

4.2.5. Comparativa de resultats i tria dels motors a entrenar

Aquest estudi es tracta d'una primera aproximació a la creació de motors de TAE per traduir guies docents i l'objectiu no és crear els motors definitius, sinó concretar els passos a seguir per entrenar-los i detectar i abordar els problemes que comporta el procés, a més de concretar el nivell de qualitat a què es pot aspirar actualment. És per això que es va decidir no entrenar els motors de totes les combinacions lingüístiques ni de totes les branques del coneixement.

Pel que fa a les combinacions lingüístiques, es va concloure que era més interessant centrar-se en les combinacions CA/ES>EN, ja que tenir un motor de TAE cap a l'anglès era molt més prioritari per la UAB, perquè l'anglès és l'idioma que molts docents no dominen gaire i, per tant, seria el motor més útil. A més, la versió en anglès de les guies docents també és la més problemàtica, perquè no és gaire fidel a l'original i conté força errors, i es volia veure la qualitat dels motors que s'obtenien.

Per decidir en quines branques del coneixement era més adequat centrar-se, es van comparar els resultats dels recomptes; concretament, els resultats que es consideraven de qualitat alta: el percentatge de segments de 10 o més paraules amb un percentatge de similitud d'entre el 90 i el 110% i el percentatge de segments de menys de 10 paraules amb entre -1 i +1 paraula de diferència. Es va calcular la mitjana dels percentatges de totes les combinacions lingüístiques de cada branca del coneixement, primer distingint el percentatge dels segments llargs (a la taula següent, "Mitjana +10p") del dels segments curts ("Mitjana -10p") i, tot seguit, unificant els resultats en un sol valor ("Mitjana total").

CORPUS	% de segments de +10 paraules: entre 90% i 110%	% de segments de -10 paraules: entre -1 i +1 paraula	Mitjana +10 p	Mitjana -10 p	Mitjana total
biociencies-en-ca	46,31%	82,03%			
biociencies-en-es	49,56%	81,41%	54,94%	85,49%	70,22%
biociencies-es-ca	68,96%	93,05%			
ciencies-en-ca	44,70%	80,61%			
ciencies-en-es	47,41%	79,91%	53,89%	**84,66%**	**69,27%**
ciencies-es-ca	69,56%	93,45%			
enginyeries-en-ca	46,12%	83,17%			
enginyeries-en-es	50,36%	81,57%	55,46%	86,10%	70,78%
enginyeries-es-ca	69,91%	93,54%			
humanitats-en-ca	55,95%	86,20%			
humanitats-en-es	57,31%	86,80%	**63,73%**	**89,39%**	**76,56%**
humanitats-es-ca	77,94%	95,17%			
salut-en-ca	44,89%	82,17%			
salut-en-es	45,31%	81,56%	**53,05%**	85,63%	69,34%
salut-es-ca	68,96%	93,16%			
socials-en-ca	50,03%	83,49%			
socials-en-es	50,44%	83,28%	57,85%	87,03%	72,44%
socials-es-ca	73,08%	94,33%			

Taula 1. Resum de resultats del recompte per branques del coneixement.

◊ Arxiu als annexos: Corpus per àmbits, recomptes i resultats > Resum resultats.xlsx

Com es pot observar a la taula, l'àmbit amb els millors resultats era el d'humanitats i el que tenia els pitjors era el de ciències. Tot i això, finalment es va decidir entrenar el de ciències de la salut en lloc del de ciències, perquè hi havia molt poca diferència entre l'un i l'altre i, a més, perquè el de ciències de la salut comptava amb, aproximadament, 50 000 segments, mentre que el de ciències en tenia 30 000. Com més extens és un corpus millor i, a part, també s'aproximava més al d'humanitats, que en tenia més o menys 100 000.

4.2.6. Preparació dels corpus definitius

Abans que res, es va considerar necessari portar a terme un últim procés de neteja dels corpus, ja que, tot i que ja se n'havien portat a terme uns quants, es va observar que hi havia molts segments que eren exclusivament cites bibliogràfiques de material de referència de les assignatures. Com que normalment la bibliografia es trobava als segments de menys de 10 paraules i les cites gairebé sempre contenien un any, es va decidir suprimir tots els segments de menys de 10 paraules que contenien alguna xifra. Això es va fer amb Notepad++, seguint aquests passos:

1. A la columna del nombre de paraules del segment original, filtrar els resultats perquè només s'hi incloguin els segments de fins a 9 paraules i seleccionar-los.

2. Menú Inici de l'Excel > Buscar y seleccionar > Ir a especial... > Solo celdas visibles (2) > Aceptar.
3. Copiar la selecció.
4. A Notepad++, Buscar > Marcar... > introduir "[0-9]" al quadre i marcar l'opció "Marcar línia" > Buscar todo.
5. Buscar > Marca > Borrar líneas marcades.

Els arxius finals poden consultar-se als annexos.

◊ Arxius als annexos: Corpus finals > [Corpus humanitats] > [EN-CA] > corpus-unic-[humanitats-en-ca]_ recompte sense bibliografia.xlsx
* Cal triar l'opció que convingui en cada cas per als noms entre claudàtors.

A partir d'aquí, calia preparar el corpus d'entrenament i el d'optimització definitius per a l'entrenament de dos motors: el d'humanitats i el de ciències de la salut.

Es va decidir que el corpus d'optimització havia d'incloure tant segments de les guies docents com dels corpus de resultats i de competències. A més, en el cas de les guies docents, es volia que les equivalències seleccionades fossin de qualitat i proporcionals a la distribució de segments per paraules.

A continuació, es mostren les taules que resumeixen la proporció de segments de cada tipus de contingut que calia incloure al corpus d'optimització ("Segments corpus optimització"). També hi ha una columna amb la quantitat de línies que hi havia d'haver al corpus d'entrenament després de suprimir els segments del corpus d'optimització ("Segments corpus entrenament").

	Segments	**% sobre total**	**Segments corpus optimització**	**Segments corpus entrenament**
Corpus únic de ciències de la salut CA-EN	45784	58,96%	590	45194
Corpus de resultats d'aprenentatge	27281	35,13%	351	26930
Corpus de competències	4587	5,91%	59	4528
TOTAL	77652		1000	

Taula 2. Taula resum de proporcions per al corpus d'optimització (ciències de la salut CA-EN).

	Segments	**% sobre total**	**Segments corpus optimització**	**Segments corpus entrenament**
Corpus únic de ciències de la salut ES-EN	44622	58,34%	583	44039
Corpus de resultats d'aprenentatge	27281	35,67%	357	26924
Corpus de competències	4587	6,00%	60	4527
TOTAL	76490		1000	

Taula 3. Taula resum de proporcions per al corpus d'optimització (ciències de la salut ES-EN).

	Segments	% sobre total	Segments corpus optimització	Segments corpus entrenament
Corpus únic d'humanitats CA-EN	66850	67,72%	677	66173
Corpus de resultats d'aprenentatge	27281	27,64%	276	27005
Corpus de competències	4587	4,65%	47	4540
TOTAL	98718		1000	

Taula 4. Taula resum de proporcions per al corpus d'optimització (humanitats CA-EN).

	Segments	% sobre total	Segments corpus optimització	Segments corpus entrenament
Corpus únic d'humanitats ES-EN	66690	67,67%	677	66013
Corpus de resultats d'aprenentatge	27281	27,68%	277	27004
Corpus de competències	4587	4,65%	46	4541
TOTAL	98558		1000	

Taula 5. Taula resum de proporcions per al corpus d'optimització (humanitats ES-EN).

◊ Ubicació de les figures als annexos: Corpus finals > [Corpus humanitats] > [EN-CA] > corpus-[salut-en-es] DEFINITIUS.xlsx (full de càlcul Resum)

* Cal triar l'opció que convingui en cada cas per als noms entre claudàtors.

L'extracció dels segments per al corpus d'optimització va consistir en diversos processos, sobretot en els corpus únics. Com ja s'ha comentat, calia separar un nombre concret de segments, que havien de ser de qualitat i representatius. Així doncs, al full de càlcul de recompte de cada branca del coneixement, en què hi havia la distribució dels segments per paraules, es va calcular per grups de 10, aproximadament, el percentatge que representaven sobre el total. Per exemple, quin percentatge representava la suma dels segments d'entre 0 i 9 paraules sobre el total de línies de tot el corpus de salut o d'humanitats. A continuació i segons el percentatge obtingut, es va calcular quants segments del corpus d'optimització corresponien a cada grup.

Distribució dels segments per paraules:		Percentatge	Percentatge	Segments sobre 677
1	2926	4,4%		
2	6220	9,3%		
3	5045	7,6%		
4	4822	7,2%		
5	4277	6,4%	52,11%	**353**
6	3488	5,2%		
7	2952	4,4%		
8	2611	3,9%		
9	2413	3,6%		
10	3702	5,6%		
11	3453	5,2%		
12	2786	4,2%		
13	2357	3,5%		
14	2141	3,2%	32,91%	**223**
15	1824	2,7%		
16	1762	2,6%		
17	1468	2,2%		
18	1311	2,0%		
19	1143	1,7%		

Il·lustració 7. Exemple del càlcul per grups dels segments corresponents al corpus d'optimització.

◊ Ubicació de la figura als annexos: Corpus finals > [Corpus humanitats] > [EN-CA] > corpus-unic-[humanitats-en-ca]_ recompte sense bibliografia.xlsx

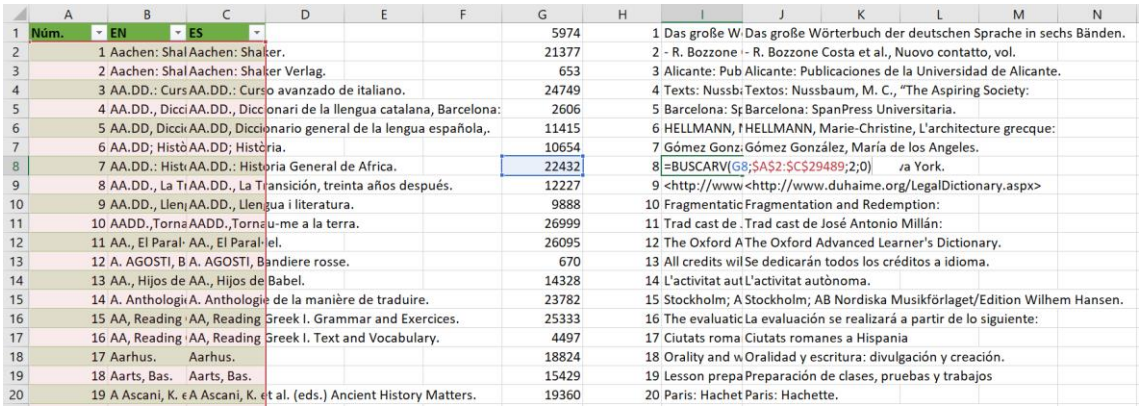
* Cal triar l'opció que convingui en cada cas per als noms entre claudàtors.

Un cop obtingut el nombre de segments que calia extreure per a cada grup, es van anar aplicant filtres a la taula per tenir els segments de qualitat (és a dir, amb una similitud d'entre el 90 i el 110% o bé d'entre -1 i +1 paraula) separats per grups de nombre de paraules. Per tant, el llibre d'Excel de cada branca del coneixement contenia un full de càlcul per a cada grup: 0-9, 10-19, 20-29, 30-39, 40-49 i 50-60+.

Com que els segments per al corpus d'optimització de cada grup no podien extreure's per ordre, es va buscar una manera de fer-ho aleatòriament mitjançant fórmules d'Excel. A continuació s'indiquen els passos que es van seguir:

1. Numerar la columna esquerra de les equivalències des de l'opció Rellenar > Series... > marcant Columnas i escrivint el nombre de línies de què disposem.
2. En una columna nova, tornar a fer una sèrie automàticament amb el nombre de segments que necessitem del grup en qüestió per al corpus d'optimització.
3. A la columna del costat, utilitzem la formula =aleatorio.entre per definir els valors entre els quals volem que Excel ens proporcioni un número aleatori. Els valors seran 1 i el nombre de línies utilitzat prèviament.

4. Copiar la formula anterior fins al final de la sèrie del nombre de segments que es necessiten i algunes caselles més.
5. Copiar i enganxar els valors aleatoris amb l'opció Pegar valores, perquè la cel·la ja sigui únicament el número enlloc de la formula. Cal suprimir els valors duplicats des de Datos > Quitar duplicados.
6. A la columna de la dreta, utilitzar la formula =buscarv i indicar que busqui el valor aleatori a la taula de les equivalències. Per exemple:
   =BUSCARV(J4;$A$2:$C$13070;2;0)
   El 2 indica la columna en què s'ha de buscar (quan es vol buscar la traducció cal posar-hi un 3) i el 0 que la coincidència ha de ser exacta.
7. Propagar les formules perquè busqui totes les equivalències corresponents als valors que necessitem per al corpus d'optimització.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Núm.	EN	ES				5974	1	Das große W	Das große Wörterbuch der deutschen Sprache in sechs Bänden.				
2	1	Aachen: Shal	Aachen: Shalter.				21377	2	- R. Bozzone	- R. Bozzone Costa et al., Nuovo contatto, vol.				
3	2	Aachen: Shal	Aachen: Shalter Verlag.				653	3	Alicante: Pub	Alicante: Publicaciones de la Universidad de Alicante.				
4	3	AA.DD.: Curs	AA.DD.: Curso avanzado de italiano.				24749	4	Texts: Nussb	Textos: Nussbaum, M. C., "The Aspiring Society:				
5	4	AA.DD., Dicci	AA.DD., Diccionari de la llengua catalana, Barcelona:				2606	5	Barcelona: Sp	Barcelona: SpanPress Universitaria.				
6	5	AA.DD, Dicci	AA.DD, Diccionario general de la lengua española,.				11415	6	HELLMANN, I	HELLMANN, Marie-Christine, L'architecture grecque:				
7	6	AA.DD; Histò	AA.DD; Història.				10654	7	Gómez Gonz	Gómez González, María de los Angeles.				
8	7	AA.DD.: Hist	AA.DD.: Historia General de Africa.				22432	8	=BUSCARV(G8	;$A$2:$C$29489;2;0)	/a York.			
9	8	AA.DD., La T	AA.DD., La Transición, treinta años después.				12227	9	<http://www	<http://www.duhaime.org/LegalDictionary.aspx>				
10	9	AA.DD., Llen	AA.DD., Llengua i literatura.				9888	10	Fragmentatic	Fragmentation and Redemption:				
11	10	AADD.,Torna	AADD.,Torna-u-me a la terra.				26999	11	Trad cast de	Trad cast de José Antonio Millán:				
12	11	AA., El Paral·	AA., El Paral·lel.				26095	12	The Oxford A	The Oxford Advanced Learner's Dictionary.				
13	12	A. AGOSTI, B	A. AGOSTI, Bandiere rosse.				670	13	All credits wil	Se dedicarán todos los créditos a idioma.				
14	13	AA., Hijos de	AA., Hijos de Babel.				14328	14	L'activitat aut	L'activitat autònoma.				
15	14	A. Anthologi	A. Anthologie de la manière de traduire.				23782	15	Stockholm; A	Stockholm; AB Nordiska Musikförlaget/Edition Wilhem Hansen.				
16	15	AA, Reading	AA, Reading Greek I. Grammar and Exercices.				25333	16	The evaluatic	La evaluación se realizará a partir de lo siguiente:				
17	16	AA, Reading	AA, Reading Greek I. Text and Vocabulary.				4497	17	Ciutats roma	Ciutats romanes a Hispania				
18	17	Aarhus.	Aarhus.				18824	18	Orality and w	Oralidad y escritura: divulgación y creación.				
19	18	Aarts, Bas.	Aarts, Bas.				15429	19	Lesson prepa	Preparación de clases, pruebas y trabajos				
20	19	A Ascani, K. e	A Ascani, K. et al. (eds.) Ancient History Matters.				19360	20	Paris: Hachet	Paris: Hachette.				

Il·lustració 8. Exemple de la mostra aleatòria per obtenir els segments del corpus d'optimització.

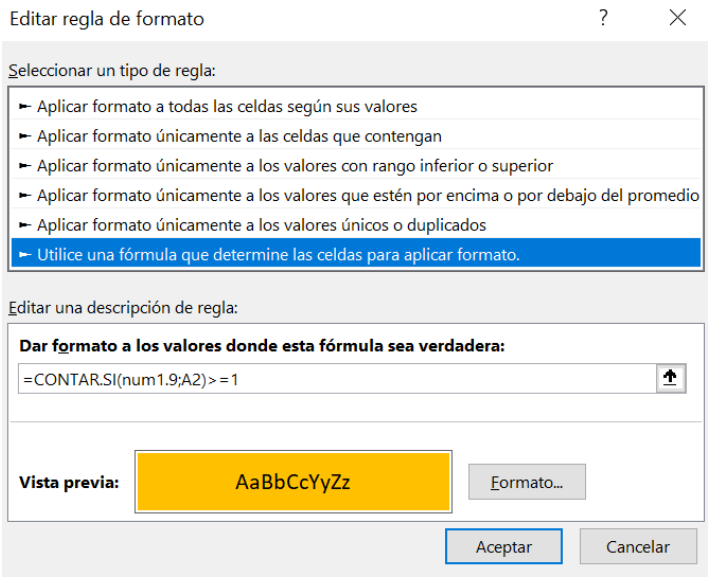
◊ Ubicació de la figura als annexos: Corpus finals > [Corpus humanitats] > [EN-CA] > corpus-unic-[humanitats-en-ca]_ recompte sense bibliografia.xlsx

* Cal triar l'opció que convingui en cada cas per als noms entre claudàtors.

Un cop extreta la mostra aleatòria, calia suprimir els segments del corpus. Això es va portar a terme mitjançant formules de l'Excel per ressaltar valors duplicats, de la manera següent:

1. Assignar un nom als números dels segments extrets (per exemple, a la columna G de la figura 12): seleccionar tota la columna > Fórmulas > Asignar nombre.
2. Seleccionar la columna on hi ha tots els números (columna A de la figura 12) > Inicio > Formato condicional > Nueva regla > Utilice una fórmula que determine las celdas para aplicar formato.
3. Utilitzar =contar.si indicant el conjunt de valors que s'han de comparar (el nom que s'ha assignat als números dels segments extrets) i el criteri que s'ha de fer servir: igual o més

gran que 1. D'aquesta manera, Excel compara els valors de les dues taules i si un apareix una vegada o més el ressalta.



Il·lustració 9. Exemple de regla per ressaltar els valors duplicats.

D'aquesta manera, es podien filtrar els resultats de la taula per color de la cel·la. Es va crear un full nou amb els segments que pertanyien al corpus d'optimització (segments ressaltats) i un altre amb els del corpus d'entrenament (segments sense emplenament). En aquest últim, també s'hi van afegir tots els segments que no complien els criteris de qualitat i que, per tant, no s'havien inclòs en els grups de què s'havien extret els segments per al corpus d'optimització.

Es va seguir el mateix procés exposat per al corpus de resultats d'aprenentatge i el de competències, però sense tenir en compte les agrupacions per nombre de paraules.

Després d'aquests processos, es van unir els corpus resultants i es va obtenir el següent:

- Corpus d'entrenament amb el corpus únic, el de competències i el de resultats d'aprenentatge, però sense els segments inclosos al corpus d'opimització.
- Corpus d'optimització amb la part proporcional de segments de cada tipus de corpus.

◊ Arxius als annexos: Corpus finals > [Corpus humanitats] > [EN-CA] > corpus-[salut-en-es] DEFINITIUS.xlsx (full de càlcul Resum)

* Cal triar l'opció que convingui en cada cas per als noms entre claudàtors.

4.3. Entrenament del motor

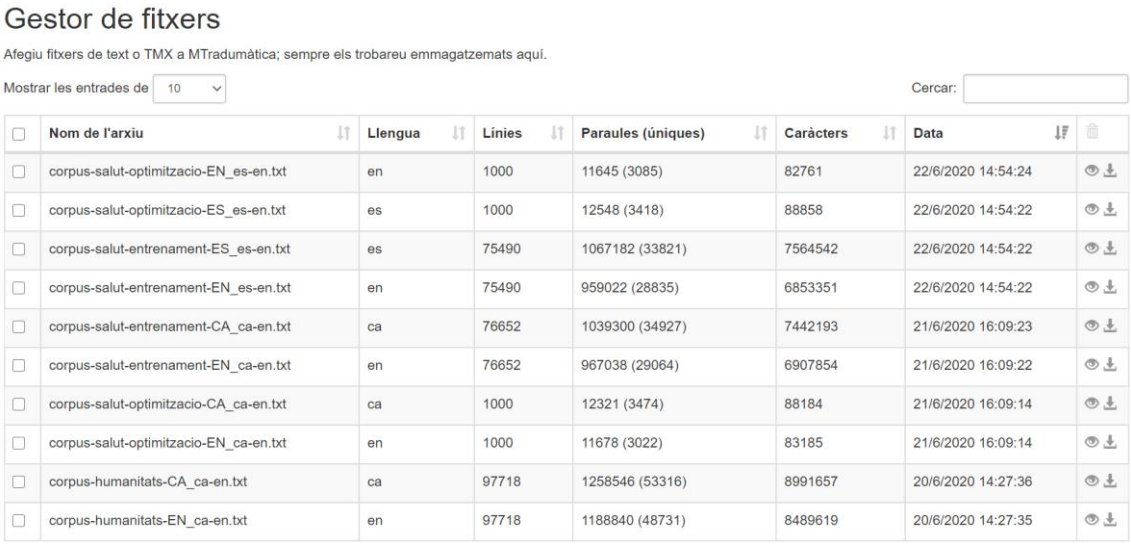
Quan es van haver creat els corpus, es va procedir a entrenar els motors. Es va utilitzar la versió local de MTradumàtica i, per tant, es va configurar VirtualBox i s'hi va importar l'arxiu

mtradumatica.ova. Un cop es van haver definit les opcions de configuració del sistema pertinents, es va procedir a entrenar els motors.

A continuació s'explica pas per pas el procés, a partir de la teoria explicada a l'apartat 3.5.

1. Carregar fitxers

Primer de tot, es va carregar el corpus d'entrenament i el corpus d'optimització de cada combinació d'idiomes a l'apartat Dades > Fitxers de MTradumàtica. Tal com es mostra a la imatge següent, es van penjar un total de 10 arxius, cadascun amb un nom molt concret perquè fos fàcil identificar-los. S'hi va fer constar la branca del coneixement de què formaven part, el tipus de corpus (entrenament o optimització), l'idioma de l'arxiu i la combinació lingüística de la guia original.


Gestor de fitxers

Afegiu fitxers de text o TMX a MTradumàtica; sempre els trobareu emmagatzemats aquí.

Mostrar les entrades de 10 Cercar:

Nom de l'arxiu	Llengua	Línies	Paraules (úniques)	Caràcters	Data
corpus-salut-optimitzacio-EN_es-en.txt	en	1000	11645 (3085)	82761	22/6/2020 14:54:24
corpus-salut-optimitzacio-ES_es-en.txt	es	1000	12548 (3418)	88858	22/6/2020 14:54:22
corpus-salut-entrenament-ES_es-en.txt	es	75490	1067182 (33821)	7564542	22/6/2020 14:54:22
corpus-salut-entrenament-EN_es-en.txt	en	75490	959022 (28835)	6853351	22/6/2020 14:54:22
corpus-salut-entrenament-CA_ca-en.txt	ca	76652	1039300 (34927)	7442193	21/6/2020 16:09:23
corpus-salut-entrenament-EN_ca-en.txt	en	76652	967038 (29064)	6907854	21/6/2020 16:09:22
corpus-salut-optimitzacio-CA_ca-en.txt	ca	1000	12321 (3474)	88184	21/6/2020 16:09:14
corpus-salut-optimitzacio-EN_ca-en.txt	en	1000	11678 (3022)	83185	21/6/2020 16:09:14
corpus-humanitats-CA_ca-en.txt	ca	97718	1258546 (53316)	8991657	20/6/2020 14:27:36
corpus-humanitats-EN_ca-en.txt	en	97718	1188840 (48731)	8489619	20/6/2020 14:27:35

Il·lustració 10. Gestor de fitxers de MTradumàtica amb els arxius penjats.

◊ Arxius als annexos: Corpus finals > [Corpus humanitats] > [EN-CA] > MTradumatica

* Cal triar l'opció que convingui en cada cas per als noms entre claudàtors.

2. Crear monotextos

Després es van crear els monotextos en anglès de cada corpus i combinació lingüística. Com que es van entrenar les combinacions CA/ES > EN, només calia entrenar models de llengua en anglès i, per tant, també només es necessitaven els monotextos en anglès. Primer s'havia de crear un monotext, posar-li nom i, a continuació, seleccionar els arxius que s'hi volien afegir.

Il·lustració 11. Gestor de monotextos de MTradumàtica amb els monotextos creats.

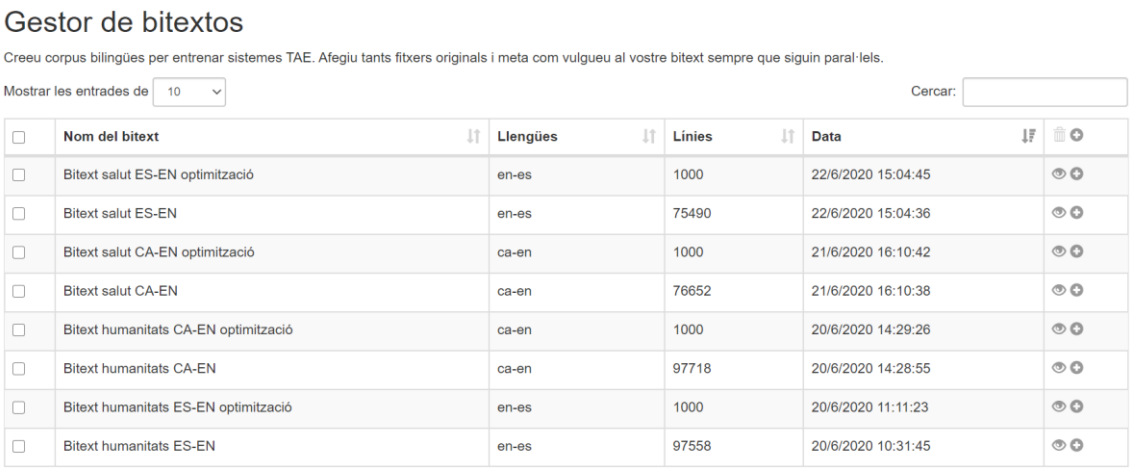
3. Entrenar models de llengua

Un cop els monotextos van estar creats, ja es podien entrenar els models de llengua. Com s'ha comentat, es van entrenar 4 models de llengua en anglès, un per a cada combinació: es va crear cada model, s'hi va donar nom i es va seleccionar el monotext corresponent. Tal com es mostra a la imatge, es van tardar pocs segons a entrenar cada model de llengua.

Il·lustració 12. Entrenador de models de llengua de MTradumàtica amb els ML creats.

4. Crear bitextos

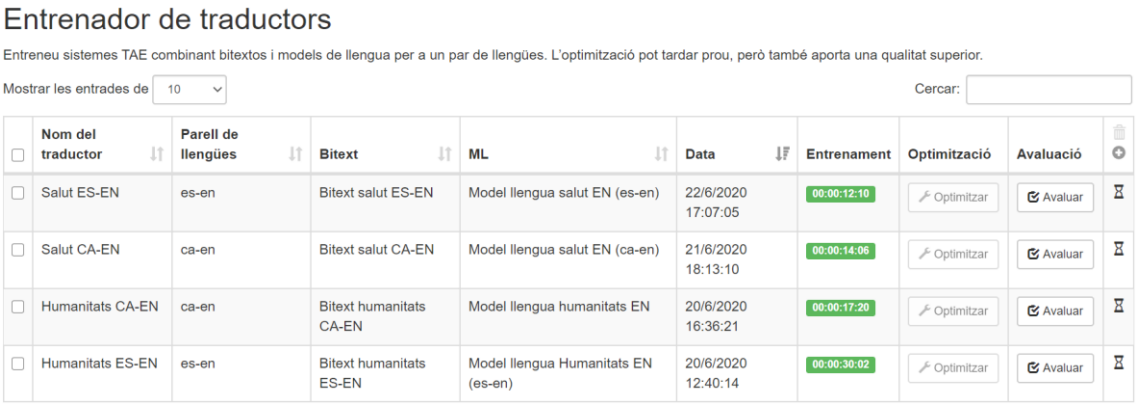
A continuació, es van crear els bitextos de cada combinació lingüística per poder entrenar els traductors automàtics. Es va seguir el mateix procés que amb els monotextos: primer es van crear i s'hi va donar nom i, tot seguit, es van seleccionar els arxius que s'hi volien afegir. Es van crear 8 bitextos, 4 per a l'entrenament i 4 per a l'optimització.



Il·lustració 13. Gestor de bitextos de MTradumàtica amb els bitextos creats.

5. Entrenar traductors

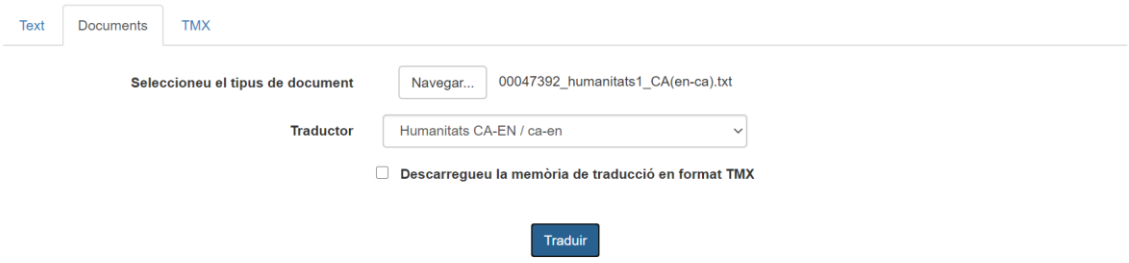
Finalment, entrenar els traductors va ser el procés més llarg. Es va posar un nom a cada traductor automàtic, definir a partir de quin bitext i quin monotext es volia fer l'entrenament i esperar entre 12 i 30 minuts que es portés a terme. Tot seguit, es va fer l'optimització de cada traductor amb el bitext corresponent, un procés que va tardar aproximadament 1 h 30 min en cada cas.



Il·lustració 14. Entrenador de traductors de MTradumàtica amb els traductors automàtics creats.

6. Traduir

Un cop finalitzat l'entrenament, es van traduir les guies docents reservades per a l'avaluació de qualitat amb l'apartat de MTradumàtica per traduir documents sencers. A l'apartat següent s'explica més detalladament.



Il·lustració 15. Secció per traduir documents de MTradumàtica.

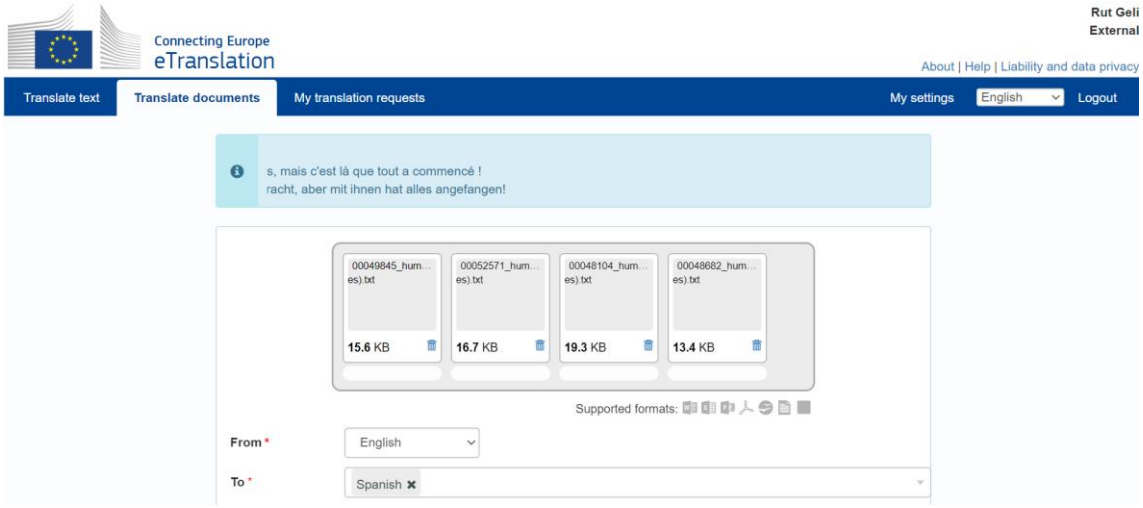
4.4. Avaluació de qualitat

Com ja s'ha comentat, va decidir portar a terme l'avaluació de qualitat amb la mètrica BLEU. Tot i que MTradumàtica té una opció per avaluar una traducció automàtica a partir d'una traducció humana, només dona una puntuació general en lloc d'una puntuació per segment. Així doncs, es va optar per l'aplicació de TILDE anomenada Interactive BLEU score evaluator, de què sí que es pot obtenir una puntuació per segment.

Com ja s'ha comentat, s'havien separat cinc guies de cada tipus per fer l'avaluació de qualitat. A partir dels arxius d'humanitats i de ciències de la salut en la combinació ES/CA>EN de cadascuna d'aquestes guies, es van crear TXT monolingües amb l'original en català o castellà o bé amb la traducció del docent en anglès. Com s'ha explicat en l'apartat anterior, es va fer la traducció de cada arxiu original amb el motor creat amb MTradumàtica corresponent.

A més, per poder fer la comparativa amb un segon motor de TA, es van traduir les guies en espanyol amb la plataforma eTranslation de la UE i, com que eTranslation no funciona amb el català, les guies en català amb Google Translate.

Per obtenir les traduccions d'eTranslation, cal tenir un compte de EU-Login i, a continuació, crear un perfil d'eTranslation. Es va indicar la llengua original i la de destí, es van penjar els arxius en TXT a la plataforma i, tot i que pot ser que es tardi 24h, la traducció va arribar al correu indicat gairebé a l'instant i en el mateix format que l'arxiu original.



Il·lustració 16. Imatge de la secció per traduir documents de la plataforma eTranslation.

Pel que fa a la TA de Google Translate, es van copiar els originals en català en un full de càlcul de Google i es va escriure la formula següent a la columna del costat:

=googletranslate(A2;"ca";"en")
*A2 és la ubicació de la cel·la que es vol traduir.

Es van crear TXT nous amb les traduccions de Google.

Quan es van tenir tots els documents a punt, es van anar introduint a TILDE. Calia incloure els arxius següents de cadascuna de les guies que es volien avaluar:

- Fitxer original (opcional).
- Fitxer amb la traducció del professor.
- Fitxer amb la traducció de MTradumàtica.
- Fitxer amb la traducció de Google Translate o bé d'eTranslation (en funció de si era una guia en català o en castellà).

TILDE dona l'opció de baixar un arxiu CSV que inclou el segment original, la traducció humana, les dues TA i la puntuació que ha donat a cadascuna. Es va baixar aquest arxiu de cadascuna de les 20 guies i es van unificar en un sol full de càlcul, en què també es van incloure altres columnes amb l'idioma i el codi de la guia original, el corpus en què estava inclosa i el nombre de paraules del segment original. D'aquesta manera, es podien filtrar els resultats i això en facilitava l'anàlisi.

◊ Arxiu als annexos: Arxius BLEU > Resultats BLEU definitiu.xlsx

Abans que res, es van observar els resultats i es van portar a terme alguns canvis, que s'expliquen a continuació.

Hi havia cel·les en blanc causades per males alineacions: o bé l'original estava en blanc o bé no hi havia la traducció de la guia docent. Així doncs, es van suprimir les files on es trobaven aquestes cel·les: Inicio > Buscar y seleccionar > Ir a especial... > Celdas en blanco > Eliminar files.

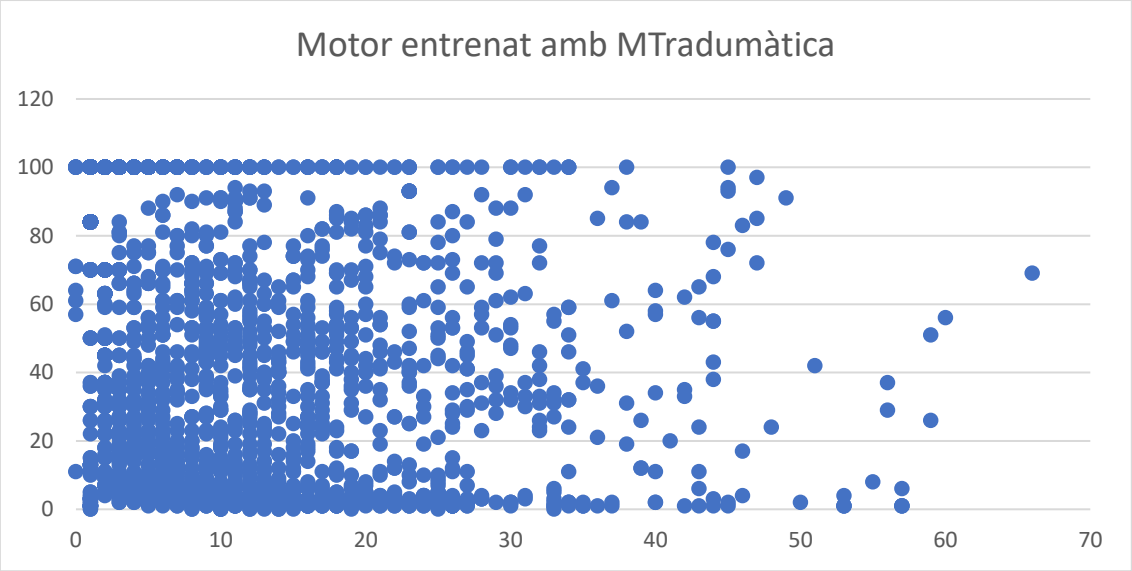
A més, es va fer una última neteja, ja que hi havia força bibliografia i segments que només contenien números i no interessava valorar aquest tipus de contingut. Així doncs, com que a l'Excel no es poden fer cerques amb expressions regulars, es van copiar els segments originals a Notepad++ i es van substituir totes les xifres ("[0-9]") per "abcdef". Es va obrir el TXT obtingut a l'Excel (Datos > Desde el texto/CSV) i es va enganxar la selecció al lloc de la taula on hi havia els segments anteriors. Tot seguit, es va fer una cerca amb "abcdef" i es va marcar "Buscar todos", es van seleccionar tots els resultats amb la tecla Shift i es va tancar el quadre per cercar. Finalment, amb les cel·les que contenien "abcdef" seleccionades (per tant, que abans contenien un número), es van eliminar les files del full de càlcul corresponents (Eliminar filas de hoja).

Per últim, es va crear un full de càlcul independent per a cada motor. Cal tenir en compte que les puntuacions d'eTranslation i de Google Translate es mostren a la columna Second Machine translated sentence i, per tant, els resultats es van separar en funció de l'idioma. Els resultats amb l'original en ES corresponien a eTranslation, mentre que els que tenien l'original en CA eren els de Google Translate.

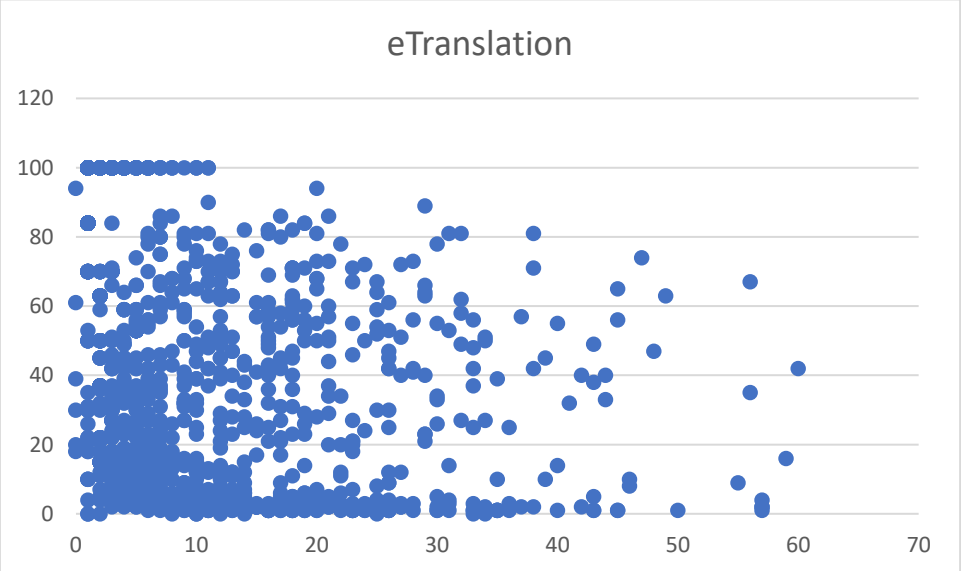
5. Resultats i anàlisi

◊ Arxiu als annexos: Arxius BLEU > Resultats BLEU definitiu.xlsx

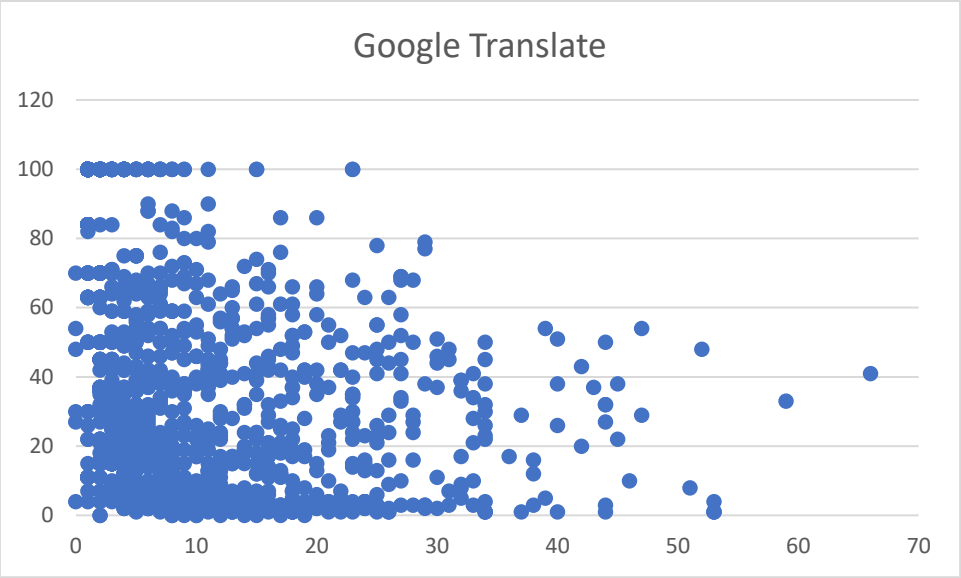
En primer lloc, es va optar per analitzar els resultats de manera visual. Així doncs, es va crear un gràfic representatiu de cada motor, que s'exposen a continuació. El nombre de paraules dels segments és el valor de l'eix de les X, mentre que la puntuació de l'avaluació BLEU es mostra a l'eix de les Y.



Il·lustració 17. Gràfic dels resultats de BLEU del motor entrenat amb MTradumàtica.



Il·lustració 18. Gràfic dels resultats de BLEU del motor eTranslation.



Il·lustració 19. Gràfic dels resultats de BLEU del motor Google Translate.

A primera vista, podem afirmar que els segments amb els millors resultats de eTranslation i Google Translate es tracten de segments molt curts, mentre que el motor entrenat amb MTradumàtica obté bons resultats tant en segments curts com en segments més llargs de fins a 30 paraules, aproximadament.

Els gràfics de eTranslation i Google Translate són força similars, però es pot apreciar un lleuger augment de la qualitat en les traduccions de eTranslation, ja que, al gràfic de Google Translate, els valors es concentren una mica més en la part inferior esquerra. Tot i això, es tracta d'una diferència gairebé imperceptible.

Tot seguit, es va decidir fer un anàlisi més exhaustiu dels millors resultats de cada motor. Per fer-ho, es van utilitzar els segments amb una puntuació mínima de 85 de BLEU, que representaven els percentatges següents sobre els segments totals corresponents:

	Segments BLEU > o = 85	**% sobre total**
MTradumàtica	1130	36,07
eTranslation	402	25,74
Google Translate	328	20,88

Taula 6. Proporció de segments amb una puntuació mínima de 85 de BLEU.

Així doncs, el motor amb més bona puntuació és l'entrenat amb MTradumàtica, seguit per eTranslation. El que presenta pitjors resultats és Google Translate.

També es va fer un recompte del nombre de segments exacte que hi havia de cada tipus juntament amb el percentatge que representava cada quantitat sobre el total.

MTradumàtica		
Núm. paraules	**Núm. segments**	**% sobre 1130**
1 a 5	882	78,05
6 a 10	125	11,06
11 a 15	44	3,89
16 a 20	30	2,65
21 a 25	16	1,42
26 a 30	10	0,88
31 a 35	8	0,71
36 a 40	3	0,27
40 a 45	3	0,27
45 a 50	6	0,53

Taula 7. Proporció de segments de 85 o més de BLEU per nombre de paraules (motor MTradumàtica).

eTranslation		
Núm. paraules	**Núm. segments**	**% sobre 402**
1 a 5	367	91,29
6 a 10	27	6,72
11 a 15	3	0,75
16 a 20	2	0,50
21 a 25	1	0,25
26 a 30	1	0,25
31 a 35	0	0,00
36 a 40	0	0,00
40 a 45	0	0,00
45 a 50	0	0,00

Taula 8. Proporció de segments de 85 o més de BLEU per nombre de paraules (motor eTranslation).

Google Translate		
Núm. paraules	**Núm. segments**	**% sobre 327**
1 a 5	295	90,21
6 a 10	24	7,34
11 a 15	4	1,22
16 a 20	2	0,61
21 a 25	2	0,61
26 a 30	0	0,00
31 a 35	0	0,00
36 a 40	0	0,00
40 a 45	0	0,00
45 a 50	0	0,00

Taula 9. Proporció de segments de 85 o més de BLEU per nombre de paraules (motor Google Translate).

Si bé és cert que la mostra de segments és més petita per a Google Translate i eTranslation, els segments de qualitat d'aquests motors són molt més curts que els de MTradumàtica i els segments de 1 a 5 paraules representen més del 90% dels segments amb més d'un 85 de BLEU. El motor entrenat amb MTradumàtica també té un bon gruix de segments d'entre 1 i 5 paraules (representen el 78%), però també en té molts d'altres amb més paraules.

Cal tenir en compte que com més curt és un segment més probable és que la traducció sigui similar a la traducció humana proporcionada, ja que hi ha menys paraules a traduir i, per tant, menys paraules en què el motor pot cometre errors. Tanmateix, això també pot ser un problema en paraules polisèmiques o casos de noms propis que poden ser considerats noms comuns, perquè hi ha menys context per discriminar la traducció correcta. S'ha cercat algun cas d'aquest tipus per observar quina traducció proporciona cada motor i s'ha observat que els tradueixen correctament, menys en algun cas concret del motor eTranslation.

Original	Traducció humana	MTradumàtica	eTranslation	Google Translate
Eunsa.	eunsa .	eunsa .	eunsa .	eunsa .
Ed. Elsevier.	ed . elsevier .	ed . elsevier .	ed . elsevier .	ed . elsevier .
Cruz.	cruz .	cruz .	oh , cruz .	cruz .

Taula 10. Exemples de traduccions de noms propis amb noms d'editorials.

També es va comparar de quin corpus provenien els segments amb millors puntuacions de BLEU, com es mostra en aquestes taules:

	MTradumàtica	% sobre 1130	eTranslation	% sobre 402	Google Translate	% sobre 327
Humanitats	625	55,31	207	51,49	139	42,51
Salut	505	44,69	195	48,51	188	57,49

Taula 11. Proporció de segments de 85 de BLEU o més pel que fa al corpus d'origen.

Tot i que en tots tres corpus els millors resultats estan força dividits pel que fa al corpus de procedència, MTradumàtica i eTranslation inclouen més bones puntuacions en segments de la branca d'humanitats, mentre que Google Translate té més segments de qualitat del corpus de salut. Sembla que, en el cas del motor entrenat amb MTradumàtica, es compleix la premissa inicial (el corpus d'humanitats era el de més qualitat), ja que hi ha més resultats de 85 de BLEU o més

provinents del corpus d'humanitats: una diferència d'un 10% amb els del corpus de ciències de la salut.

També es va analitzar quants dels millors segments de MTradumàtica provenien d'un original en català i quants d'un original en castellà, i es va arribar a la conclusió que la qualitat era força similar:

CA	559	49,47%
ES	571	50,53%

Taula 12. Proporció de segments de 85 de BLEU o més de MTradumàtica pel que fa a l'idioma d'origen.

A part d'aquest anàlisi dels millors resultats de cada tipus de corpus i de motor, també s'ha portat a terme un anàlisi qualitatiu observant segments aleatoris i s'han arribat a conclusions interessants que s'exposen a continuació.

- ➔ Primer de tot, cal tenir en compte que la mètrica BLEU dona un resultat aproximat, que calcula la puntuació en funció de les paraules de la traducció que coincideixen amb l'original, sense tenir en compte l'ordre de la frase ni els sinònims i marcant com a errors canvis de majúscules i minúscules, per exemple.

- ➔ Un altre aspecte a comentar és que cap traducció, inclosa la humana, comença amb majúscula i hi ha espais entre les paraules i els signes de puntuació. Això és degut a la tokenització i al *truecasing* automàtics de TILDE, però crida l'atenció que també han perdut les majúscules els noms propis, que normalment es mantenen en aquest tipus de processos.

- ➔ S'ha observat que, en la majoria dels casos, tots tres motors tradueixen correctament els segments que conten exclusivament números i anys. Pel que fa a les cites bibliogràfiques, MTradumàtica acostuma a reconèixer-les, mentre que Google Translate i eTranslation solen traduir-les. Per exemple:

Original	-Pujol, Josep, La memòria literària de Joanot Martorell:
Traducció humana	pujol , josep , la memòria literària de joanot martorell :
MTradumàtica	- pujol , josep , la memòria literària de joanot martorell :
Google Translate	- pujol , joseph eisner literary memory :

Taula 13. Exemple de traducció de cites bibliogràfiques (1).

Hi ha algunes cites bibliogràfiques que durant el procés d'alineació es van separar de la resta de la cita i per aquest motiu no són fàcilment detectables com a cites. Tot i això, MTradumàtica les continua mantenint en l'idioma original:

Original	Del carácter al contexto:
Traducció humana	del carácter al contexto :
MTradumàtica	del carácter al contexto :
eTranslation	from character to context :

Taula 14. Exemple de traducció de cites bibliogràfiques (2).

➔ També cal recordar que les traduccions humanes en base a què BLEU ha obtingut les puntuacions no són de gaire bona qualitat i això fa que, en alguns casos, no siguin la millor traducció possible, fet que condiciona la puntuació del BLEU.

Hi ha situacions en què es produeixen omissions, com ara el següent:

Original	Lectura de textos i treball de materials audiovisuals.
Traducció humana	reading texts and audiovisual materials .
MTradumàtica	reading texts and audiovisual materials .
Google Translate	reading texts and audiovisual materials work .

Taula 15. Exemple d'error en la traducció humana (1).

Però també hi ha casos més greus, en què la traducció humana és errònia. Generalment, això es deu a dos motius: que no s'hagi traduït bé (1) o que formés part d'una llista i s'hagi desordenat (2).

(1) Pot ser que el nivell de llengua anglesa del docent que ha traduït la guia no sigui òptim i cometi errors. Per exemple, a l'hora de traduir «equip docent», sovint s'empra «teachers» quan la traducció correcta és «teaching staff». Quan Google Translate o eTranslation utilitzen aquesta forma, BLEU els penalitza, perquè no és la paraula que hi ha a la traducció humana.

(2) També hi ha casos en què la traducció humana és la traducció d'un altre segment. Es tracten de segments que no van suprimir els processos de neteja que, a la guia docent, formaven part d'una llista que el professor va traduir de manera desordenada. Per exemple:

Original	Produir textos escrits en llengua A per poder traduir.
Traducció humana	solving translation problems from different specialisation fields (legal , financial , scientific , technical , literary , audiovisual texts , localization) .
MTradumàtica	producing written texts in language a in order to translate .
Google Translate	produce written texts in order to translate .

Taula 16. Exemple d'error en la traducció humana (2).

Tots aquests errors porten MTradumàtica a fer traduccions errònies, ja que està entrenat amb aquests corpus, i, a més, també afecten l'avaluació.

➔ Un altre aspecte que s'ha observat és que el motor entrenat amb MTradumàtica, com que s'ha creat a partir de les guies docents, utilitza una terminologia molt similar a la de les traduccions dels professors amb què s'ha comparat, cosa que és un aspecte molt interessant. Per exemple, a l'exemple següent, MTradumàtica utilitza «didactic» i «autonomy» en lloc d'altres sinònims menys adequats, com fa Google Translate («educational» i «independence»):

Original	Les activitats didàctiques s'organitzen en tres blocs, segons el grau d'autonomia requerida per part de l'estudiant:
Traducció humana	the didactic activities are organized in three blocks , according to the degree of autonomy required by the student :
MTradumàtica	the didactic activities are organized in three blocks , according to the degree of autonomy required by the student :
Google Translate	the educational activities are organized into three groups according to the degree of independence required by the student :

Taula 17. Exemple sobre la terminologia de les guies.

També cal dir que aquest fet influeix en les puntuacions, ja que, com s'ha comentat, BLEU no té en compte els sinònims i utilitzar els mateixos termes que la traducció humana suposa un avantatge respecte als altres motors.

➔ Tot i que MTradumàtica està entrenat amb algunes traduccions errònies, en alguns casos també és capaç de millorar la traducció humana proposada, probablement gràcies a la resta de corpus:

Original	Antecedents (definició del problema, epidemiologia, factors associats, estat actual del tema).
Traducció humana	antecedents (definition of the problem , epidemiology , associated factors , current state of the subject) .
MTradumàtica	background (definition of the problem , epidemiology , associated factors , current state of the subject) .
Google Translate	background (problem definition , epidemiology , factors associated current state of the subject) .

Taula 18. Exemple de millora de la traducció humana.

➔ MTradumàtica, a diferència dels altres motors de TA, no tradueix correctament els URL, com es mostra a continuació:

Original	http://www.language.berkeley.edu/fanjian/toc.html (correspondencias)
Traducció humana	http://www.language.berkeley.edu/fanjian/toc.html (correspondències)
MTradumàtica	http : \ / \\ / www.language.berkeley.edu / fanjian / toc.html (correspondències)
eTranslation	http://www.language.berkeley.edu/fanjian/toc.html (correspondences)

Taula 19. Exemple de traducció d'URL (1).

Tot i això, aquest error podria solucionar-se amb una substitució posterior a la traducció automàtica. Concretament, caldria fer-ne dues:

- "\ /" per "/", que resol la majoria de casos:

http://lost - theory.org/ocrat/c	http : \ / \ / lost - theory.org / ocrat / chargif \ /
https://www.pleco.com/	https : \ / \ / www.pleco.com \ /
http://lost - theory.org/ocrat/c	http : \ / \ / lost - theory.org \ / ocrat \ / chargif \ /

Taula 20. Exemple de traducció d'URL (2).

- "\\ /" per "/", per a alguns casos concrets com aquests:

http://www.aeped.	http://www.aeped.es/protoco http : \ / \\ / www.aeped.es / protocolos / index.htn
http://www.aeped.	http://www.aeped.es/protoco http : / / www.aeped.es \\ / protocols \ / index.htm

Taula 21. Exemple de traducció d'URL (3).

➔ Per últim, i concretament sobre els resultats de MTradumàtica que són bons però no perfectes, d'entre 85 i 99 de BLEU, s'ha observat que, en general, la qualitat és bona, ja que les traduccions segueixen bastant les traduccions humanes, però que tenen algunes mancances. S'han observat tres problemàtiques, que es mostren en tres exemples representatius: s'omet certa informació (1), s'inclouen símbols estranys totalment inesperats (2) i es deixa alguna paraula en l'idioma original (3).

(1) 91 de BLEU

Original	- Es absolutamente indispensable que tanto la expresión oral como la escrita sean correctas en el contenido y en la forma, tal y como ha de ser exigible a quien haya superado unos estudios de bachillerato y aspire a un título universitario, y sea cual sea la lengua utilizada.
Traducció humana	- it is absolutely essential that both oral and written expression are correct in content and form , as must be required of anyone who has completed high school and aspires to a university degree , regardless of the language used .
MTradumàtica	- it is absolutely essential that both oral and written expression are correct in content and form , as must be required of anyone who has completed high school and aspires to a university degree , and the language used .
eTranslation	— it is absolutely essential that both oral and written expression be correct in the content and form , as must be required of anyone who has completed a baccalaureate degree and aspires to a university degree , and whatever language is used .

Taula 22. Exemple d'omissió d'informació en la traducció de MTradumàtica.

(2) 85 de BLEU

Original	- És absolutament indispensable que tant l'expressió oral com l'escrita siguin correctes en el contingut i en la forma, tal i com ha de ser exigible a qui hagi superat uns estudis de batxillerat i aspiri a un títol universitari, i sigui quina sigui la llengua utilitzada.

Traducció humana	- it is absolutely essential that both oral and written expression are correct in content and form , as must be required of anyone who has completed high school and aspires to a university degree , regardless of the language used .
MTradumàtica	- it is absolutely essential that both ’ ’ oral and written expression are correct in content and form , as must be required of anyone who has completed high school and aspires to a university degree , and the language used .
Google Translate	- it is absolutely essential that both oral and written are correct in content and form , as who should be required to have completed high school studies and aspires to a college degree , and whatever the language used .

Taula 23. Exemple de símbols estranys en la traducció de MTradumàtica.

(3) 86 de BLEU

Original	- Las tutorías se reservan para dudas específicas y particulares, dejando las de interés colectivo para lasintervenciones en el aula.
Traducció humana	- tutorials are reserved for specific and particular doubts , leaving those of collective interest for interventions in the classroom .
MTradumàtica	- tutorials are reserved for specific and particular doubts , leaving those of collective interest for lasintervenciones in the classroom .
eTranslation	— tutorials are reserved for specific and particular doubts , leaving those of collective interest for interventions in the classroom .

Taula 24. Exemple de paraules en l'idioma original en la traducció de MTradumàtica.

6. Conclusions

Com es comentava a la introducció d'aquest treball, la TA és una eina molt útil en molts àmbits de la traducció i podem concloure que un d'aquests àmbits és la traducció de guies docents d'assignatures. Tot i que el motor que s'ha entrenat amb MTradumàtica té diverses mancances que cal solucionar, també s'ha observat que té potencial per esdevenir un bon motor de TAE.

Tot i que l'avaluació feta amb BLEU té certa predisposició a puntuar millor MTradumàtica que els altres motors perquè és més similar a les traduccions humanes fetes servir de referència, la qualitat de les traduccions del motor obtingut són de bona qualitat pel que fa als segments amb millors puntuacions. Si bé és cert que s'han observat alguns errors freqüents, la majoria podrien solucionar-se amb una millora del corpus d'entrenament: aplicant-hi encara més processos de neteja o, idealment, millorant-ne les traduccions. També caldria cercar la manera de reordenar les traduccions de llistes desordenades o bé de suprimir-les.

Si aquesta millora del corpus es portés a terme, molts dels errors actuals no es cometrien (omissions, deixar paraules en l'idioma original, etc.). Això combinat amb la correcta detecció de les cites bibliogràfiques i el bon ús de la terminologia de les guies docents que ja sembla oferir el motor de MTradumàtica, faria que els resultats fossin millors que els de qualsevol altre motor de TA. Tot i això, també caldria investigar perquè aquest motor a vegades inclou símbols estranys a les traduccions.

Una altra millora que es podria aportar al projecte seria l'addició d'un corpus de terminologia, una idea que es va plantejar però que finalment no es va portar a terme, perquè l'extracció terminològica no tenia prou bons resultats amb el corpus actual de guies docents. Si es millorés el corpus, segurament també seria més fàcil fer aquesta extracció, que suposaria un valor afegit i potser serviria perquè el motor utilitzés una terminologia encara més específica i unificada.

Per acabar, també es vol posar en valor la plataforma MTradumàtica, ja que és una eina molt útil a l'hora d'entrenar motors de TA. Després de fer-la servir, es pot afirmar que la interfície és molt intuïtiva i que és fàcil d'utilitzar per a persones que no són expertes en l'entrenament de motors de TA, com és el meu cas.

7. Bibliografia

Arevalillo Doval, Juan José. «La traducción automática en las empresas de traducción». Tradumàtica, Núm. 10 (2012), p. 179-184. DOI 10.5565/rev/tradumatica.19 <https://ddd.uab.cat/record/105644> [Consulta: 19/03/2020].

Babych, Bogdan. «Automated MT evaluation metrics and their limitations». Tradumàtica, Núm. 12 (2014), p. 464-470. DOI 10.5565/rev/tradumatica.70 <https://ddd.uab.cat/record/130148> [Consulta: 13/04/2020].

Calude, Andreea. (2003). «Machine translation of various text genres». 46. https://www.researchgate.net/publication/228938192_Machine_translation_of_various_text_genres [Consulta: 13/05/2020].

Casacuberta Nolla, F.; Peris Abril, A. (2017). «Traducción automática neuronal». Revista Tradumàtica. Tecnologies de la Traducció, 15, 66-74. https://doi.org/10.5565/rev/tradumatica.203 [Consulta: 13/05/2020].

Doğru, Gökhan; Martín Mor, Adrià; Aguilar-Amat, Anna. «Parallel corpora preparation for machine translation of low-resource languages : Turkish to English cardiology corpora». A: Proceedings of the LREC 2018 Workshop 'MultilingualBIO: Multilingual Biomedical Text Processing'. 2018, p. 12-15. <https://ddd.uab.cat/record/196871> [Consulta: 9 abril 2020].

Forcada, Mikel (2009). «Apertium: traducció automàtica de codi obert per a les llengües romàniques». Recuperat de: https://www.dlsi.ua.es//~mlf/docum/forcada09j.pdf [Consulta: 08/04/2020].

Forcada, M. L. (2017). «Making sense of neural machine translation». Translation Spaces, 6:2, p. 291-309. Recuperat de: https://www.dlsi.ua.es/~mlf/docum/forcada17j2.pdf [Consulta: 08/04/2020].

Hearne, Mary, i Andy Way. 2011. «Statistical Machine Translation: A Guide for Linguists and Translators: SMT for Linguists and Translators». *Language and Linguistics Compass* 5 (5): 205-26. Recuperat de: https://www.computing.dcu.ie/~away/CA446/SMTforLinguists.pdf [Consulta:10/04/2020].

Koehn, Philipp. 2010. «Statistical machine translation». Cambridge; New York: Cambridge University Press. Recuperat de: https://www.statmt.org/book/ [Consulta: 13/04/2020].

Lavie, Alon, i Denkowski, Michael (2009). «The METEOR Metric for Automatic Evaluation of Machine Translation». Recuperat de: https://www.cs.cmu.edu/afs/cs.cmu.edu/project/mteval-1/Papers/MT-Journal-2009/meteor-mtj-2009.pdf [Consulta: 15/04/2020].

Papineni, Kishore, Salim Roukos, Todd Ward, i Wei-Jing Zhu. 2002. «BLEU: a method for automatic evaluation of machine translation». En *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Association for Computational Linguistics. Recuperat de: http://dl.acm.org/citation.cfm?id=1073135 [Consulta: 13/04/2020].

Sánchez-Martínez, Felipe (2011). «Choosing the best machine translation system to translate a sentence by using only source-language information». Proceedings of the 15th Annual Conference of the European Association for Machine Translation, 97-104, Mikel L. Forcada and Heidi Depraetere and Vincent Vandeghinste, Leuven (Belgium), European Association for Machine Translation. Recuperat de: https://www.dlsi.ua.es//~fsanchez/pub/pdf/sanchez-martinez11b.pdf [Consulta: 10/04/2020]

Sánchez-Martínez, Felipe. «Motius del creixent ús de la traducció automàtica seguida de postedició.». Tradumàtica: traducció i tecnologies de la informació i la comunicació, [en línia], 2012, Núm. 10, p. 150-6. Recuperat de: https://www.raco.cat/index.php/Tradumatica/article/view/263223 [Consulta: 19/03/2020].

SYSTRAN blog (2016). «How does Neural Machine Translation work?». Recuperat de: https://blog.systransoft.com/how-does-neural-machine-translation-work/ [Consulta: 05/04/2020].

United Language Group (2016). «Making the most of machine translation today». Recuperat de: https://46axn43qhl4w23sv6y5n2p1u-wpengine.netdna-ssl.com/wp-content/uploads/2017/07/ULG_eBook_MachineTranslation.pdf [Consulta: 05/04/2020].

8. Annexos

S’adjunten els annexos del treball comprimits en un arxiu ZIP.