TREBALL DE FINAL DE MÀSTER
2022-2023

AVALUACIÓ DE LA UTILITAT DE L'EINA DE NETEJA DE CORPUS AUTOMÀTICA BICLEANER

AVALUACIÓ APLICADA AL CORPUS WIKIMATRIX ANGLÈS-CATALÀ EN COMPARACIÓ AMB L'AVALUACIÓ MANUAL DE KEOPS

MÀSTER EN TRADUMÀTICA: TECNOLOGIES DE LA TRADUCCIÓ
FACULTAT DE TRADUCCIÓ I INTERPRETACIÓ

Mar Balaguer Falcó

TUTOR
Marc Riera Irigoyen

Barcelona, 23 de juny de 2023

Dades del TFM/ Datos del TFG / *Dissertation data*

Títol: Avaluació de la utilitat de l'eina de neteja de corpus automàtica Bicleaner: avaluació aplicada al corpus WikiMatrix anglès-català en comparació a KEOPS

Título**:** Evaluación de la utilidad de la herramienta de limpieza de corpus automática Bicleaner: evaluación aplicada al corpus WikiMatrix inglés-catalán en comparación con la evaluación manual de KEOPS

Title: Evaluation of the usefulness of the automatic corpus cleaning tool Bicleaner: evaluation applied to the corpus WikiMatrix English-Catalan in comparison to the KEOPS manual evaluation

Autora: Mar Balaguer Falcó

Author: Mar Balaguer Falcó

Tutor: Marc Riera Irigoyen

Tutor: Marc Riera Irigoyen

Centre: Universitat Autònoma de Barcelona (UAB)

Centro: Universidad Autónoma de Barcelona (UAB)

Centre: Autonomous University of Barcelona (UAB)

Estudis: Màster oficial en Tradumàtica: Tecnologies de la Traducció

Estudios: Máster oficial en Tradumática: Tecnologías de la Traducción

Studies: Official master's degree in Tradumatics: Translation Technologies

Paraules clau/ Palabras clave / *Keywords*

Avaluació de corpus, Bicleaner, KEOPS, anglès-català, WikiMatrix

Evaluación de corpus, Bicleaner, KEOPS, inglés catalán, WikiMatrix

Corpus evaluation, Bicleaner, KEOPS, English-Catalan, WikiMatrix

Resum del TFM / Resumen del TFM / *Abstract*

El present treball de fi de màster té com a objectiu avaluar la qualitat de l'eina de neteja i avaluació de corpus automàtica, Bicleaner. Per aconseguir-ho s'ha passat un mostreig del corpus de WikiMatrix anglès-català de 200 feta amb segments aleatoris per l'eina. Per extreure conclusions sobre els resultats sobre aquest mateix conjunt de segments s'ha realitzat altra avaluació, de forma manual, amb ajuda de l'eina KEOPS i se n'han comparat els resultats. Els resultats mostren que l'avaluació automàtica pot ser molt útil per a fer una primera aproximació, encara que no produeixi uns resultats prou coherents per a dependre'n únicament i es requereixi una segona avaluació.

El presente trabajo de final de máster tiene como objetivo evaluar la calidad de la herramienta de limpieza y evaluación de corpus automática, Bicleaner. Para conseguirlo se ha pasado un muestreo del corpus de WikiMatrix inglés-catalán hecha con segmentos aleatorios por la herramienta. Para extraer conclusiones sobre los resultados se ha realizado otra evaluación, de forma manual, con ayuda de la herramienta KEOPS y se han comparado los resultados. Los resultados muestran que la evaluación automática puede ser muy útil para hacer una primera aproximación, aunque luego no produzca unos resultados lo suficientemente coherentes para depender de ella únicamente y se necesite una segunda evaluación.

The aim of this master's thesis is to evaluate the quality of the automatic corpus cleaning and evaluation tool, Bicleaner. To achieve this, a sample of the English-Catalan WikiMatrix corpus made with random segments has been evaluated with the tool. In order to draw conclusions about the results, another evaluation was carried out manually with the help of the KEOPS tool and the results were compared. The results show that the automatic evaluation can be very useful for a first approximation, even though it doesn't produce coherent enough results to be used on its own and a second evaluation is needed.

Agraïments

Aquest TFM no hauria estat possible sense l'ajuda del meu tutor, la meva família, la meva parella, els meus amics i els meus companys de màster. Vull agrair a Marc Riera, el meu tutor, la seva paciència inacabable amb els meus correus incomprensibles a les 23 de la nit i el seu recolzament durant tot el procés. Vull donar-li les gràcies també a ma mare, Roser, que des d'Alacant m'ha escoltat desfogar-me per telèfon cada dia, durant hores, tot i que no entenia res. També vull agrair a la meva parella, el Pere, tot el seu recolzament i la seva confiança en mi. I, per acabar, vull donar-li les gràcies també a la meva cosina Paula i al meu amic Nil, els quals em van ajudar amb els aspectes més tècnics d'Excel. Sense tots ells, no hauria pogut donar la millor versió de mi mateixa.

A més, també em vull donar les gràcies a Mar, qui s'ha passat hores i hores fent recerca a llocs perduts a Internet, qui, tot i patir d'atacs d'ansietat, seguia escrivint i treballant en aquest projecte mentre plorava. Qui no ha deixat que la depressió ni l'ansietat trunquessin la seva carrera.

ÍNDEX

ÍNDEX DE FIGURES I TAULES

1. Introducció

En un món on milions de persones cada dia fan servir els traductors automàtics per al seu dia a dia, resulta sorprenent com no són conscients de tot el que hi ha al darrere per fer-los funcionar. S'ha d'excavar una mica, però aleshores es troba el món dels corpus lingüístics, que són la base d'aquests sistemes, entre moltes altres coses. No només s'utilitzen per entrenar motors de traducció automàtics, sinó que es fan servir per estudiar llengües, per entrenar motors de processament natural del llenguatge, com a eines pedagògiques, etc. Milers de persones treballen gairebé en l'anonimat recollint textos, creant corpus, netejant aquests corpus i, avaluant-los, cosa que permet als usuaris normals i corrents poder utilitzar aquestes tecnologies que depenen dels corpus lingüístics.

Amb els avenços tecnològics que tenim actualment i gràcies a la recerca que fan els lingüistes, s'han pogut desenvolupar unes eines que faciliten la feina als lingüistes a l'hora d'enllestir corpus més ràpidament i amb major qualitat. N'hi ha de tot tipus: eines de recollida de textos, eines de neteja de corpus, eines d'avaluació de corpus, eines d'automatització de tasques, etc. Aquest TFM pretén agafar una d'aquestes eines i avaluar-ne la utilitat.

En aquest treball de fi de màster el que faré serà agafar un corpus de mala qualitat, WikiMatrix, en la combinació anglès-català i utilitzar-lo per comprovar la utilitat d'una de les eines més populars de neteja i avaluació de corpus de forma automàtica: Bicleaner. A més, per comprovar si l'avaluació que proporciona aquesta eina és correcta, contrastaré els seus resultats amb una avaluació manual feta per mi, d'una mostra de 200 segments del corpus. Per fer-ho utilitzaré l'eina KEOPS, així que també provaré com funciona aquesta eina, tot i que per a aquest treball només farà la funció d'eina assistent. Una vegada presentat el marc teòric, importantíssim per entendre què són els corpus i la traducció automàtica, d'on venen i quins usos tenen, presentaré la metodologia, on explicaré més en detall quins programes he emprat i com ha sigut el procés d'instal·lació i aplicació de les eines, a part de la preparació del corpus. Tot seguit presentaré els resultats, amb ajuda d'uns gràfics, i, finalment, n'extrauré unes conclusions.

1.1. Justificació

El meu gran interès per les tecnologies i per les llengües, i sobretot per la manera en què convergeixen contínuament i depenen les unes de les altres, m'ha fet voler investigar més en aquest món, especialment en la creació i neteja de corpus. Gràcies al màster vaig descobrir l'existència dels corpus

i les seves utilitats, i em va semblar que era un camp poc explorat. A més, una vegada endinsada en la recerca del marc teòric, vaig descobrir l'existència de les eines per a corpus, i vaig pensar que fer-ne un TFM al respecte donaria visibilitat a aquest camp. Trobo enriquidor poder aprofundir més sobre els corpus i com funcionen, ja que, per desgràcia, al màster no els hi vam dedicar tant temps com m'hagués agradat.

No només això, sinó que sempre he estat molt interessada en les llengües minoritzades, com a parlant de català, i volia posar el meu granet de sorra fent servir un corpus d'anglès-català, com ara el WikiMatrix. Aquest corpus tenia, a més, un problema afegit, i era que, segons el GitHub de Softcatalà, un 30% dels 1.509.909 segments que el conformen era de mala qualitat, estava mal alineat o tenia llengües incorrectes. Per això vaig decidir que era un molt bon corpus per provar aquestes eines de neteja i avaluació de corpus.

1.2. Objectius

L'objectiu del present treball és avaluar la utilitat de l'eina de neteja i avaluació Bicleaner. Per aconseguir-ho pretenc extreure uns 200 segments de mostra del corpus WikiMatrix anglès-català, avaluar-los amb l'eina i, posteriorment, comparar-los amb una avaluació manual, per veure com són de coherents les puntuacions del Bicleaner. Aquesta avaluació manual la realitzaré amb ajuda de l'eina KEOPS. Una vegada analitzats i comparats els resultats de totes dues eines, parlaré de si Bicleaner em sembla una bona eina de neteja i avaluació de corpus o no.

A llarg termini el que pretenc amb aquest treball de final de màster és motivar a altres alumnes i lingüistes a endinsar-se una mica més en el fascinant món dels corpus lingüístics, la seva gestió i les diferents eines que existeixen.

2. Marc teòric

2.1. Corpus

La quantitat de definicions del que són els corpus és tan gran i tan diversa que no hi ha manera de representar totes les opcions. Per això, amb l'objectiu de representar una mica com ha evolucionat aquesta definició, n'hi posaré tres, de tres anys diferents: 1996, 2006 i 2021.

Segons l'associació *Expert Advisory Group on Language Engineering Standards* (EAGLES) (1996), un corpus és una col·lecció de fragments de llenguatge que han estat seleccionats i ordenats segons uns criteris lingüístics explícits perquè es puguin utilitzar com a mostra d'aquella llengua. No obstant això, EAGLES, fa una distinció entre els corpus lingüístics i els corpus computacionals i defineix els corpus computacionals com corpus codificats i homogeneïtzats per a tasques de recuperació obertes.

L'any 2006 McEnery, Xiao i Tono van definir els corpus com un recull de textos representatius, orals o escrits, en un format llegible per ordinador que pot estar anotat amb informació lingüística diversa.

Acabaré amb una definició de corpus una mica més recent, de Rojo (2021), que diu que un corpus és un conjunt de textos o fragments de textos, orals o escrits, produïts en condicions naturals, conjuntament representatius d'una llengua o varietat lingüística que s'emmagatzemen en format electrònic i es codifiquen amb la intenció de ser analitzats científicament.

Per saber què és un corpus no només hem de tenir en compte la definició literal del terme, sinó que és interessant tenir en compte les característiques que han de complir, pel que fa al disseny.

EAGLES, va definir l'any 1996 uns criteris bàsics que s'han de complir perquè un conjunt de textos en format electrònic es consideri un corpus.

En primer lloc, un corpus havia de ser tan gran com permetés la tecnologia en el moment de la seva creació. En segon lloc, el corpus havia de tenir diverses mostres de textos de gèneres diferents per representar la llengua de forma acurada. En tercer lloc, els textos del corpus s'havien de classificar segons els seus gèneres. En quart lloc, totes les mostres havien de ser de la mateixa mida. Per últim, s'havia de declarar l'origen del corpus.

Tot i que aquests criteris es van crear amb la intenció que els corpus representessin la llengua de forma fidel, no hi ha gaires corpus actuals que els segueixin. No només perquè els paradigmes i els objectius

de la creació de corpus han anat canviant i avançant amb la tecnologia, sinó perquè altres lingüistes consideren que és més important la quantitat d'informació que hi ha a un corpus que no pas la fidelitat i representació. No obstant això, a mesura que han passat els anys i ha anat avançant la tecnologia, aquestes "normes" o característiques han anat canviant i s'han anat adaptant als nous temps.

Segons Llamazares (2008), els criteris són els següents:

1. El corpus ha d'estar informatitzat, cosa que permet cercar i localitzar informació fàcilment, recuperar informació, observar-ne la freqüència i classificar les dades segons diferents criteris.
2. Les dades han de ser autèntiques per representar correctament la llengua.
3. Els textos que s'han seleccionat per al corpus han de seguir uns criteris, lingüístics o extralingüístics, de selecció.
4. Els textos han de representar la varietat lingüística que s'estudia. Quan parlem de varietat ens referim, no només al gènere textual, sinó a l'autor, el dialecte o l'època, entre d'altres.
5. Els corpus han de tenir, en teoria, un nombre finit d'elements.

Amb l'objectiu de comparar el disseny de corpus i la seva evolució amb els anys també citaré Rojo (2021), que explica que és el que normalment constitueix un corpus. És important remarcar que no proposa unes normes rígides de com s'haurien de construir, sinó que té un enfocament més descriptiu, ja que ell mateix indica que aquestes característiques no representen tota la realitat.

1. Un dels conceptes fonamentals dels corpus és la representativitat. En general, els corpus estan constituïts per una selecció de textos existents i vinculen a les mostres amb la població, de manera que la representació sigui rellevant.
2. S'ha de tenir en compte que el corpus estigui equilibrat, és a dir, que totes les mostres tinguin aproximadament el mateix volum.
3. Els corpus haurien d'estar en format electrònic. Aquesta exigència es deriva del desenvolupament de les tecnologies.
4. S'han d'anotar els corpus, ja que com tenen tanta informació, aquesta és l'única forma de consultar posteriorment la informació.
5. Actualment no se sol tenir en compte la mida, ja que no hi ha límits gràcies a les noves tecnologies.
6. Els corpus actuals, és que els més generals (com ara el CREA o el CORPES) contenen tant mostres escrites com orals.

És important destacar, però, que en l'actualitat gairebé totes aquestes característiques són modificables, ja que cada corpus respon a un objectiu diferent i pretén aconseguir coses diferents. Un corpus pot ser total o de mostres, obert o tancat, general o especialitzat, multimodal, monolingües o multilingües, etc.

Com es pot observar, a més, el que ve a ser "l'esquelet" del que es considera un corpus, com ara la representativitat o l'equilibri, entre d'altres, s'ha mantingut amb el temps.

2.1.1. Història dels corpus

La història dels corpus és una mica vaga i confusa, ja que els lingüistes no es posen d'acord sobre quan es considera que comença, exactament. Actualment, una de les característiques que defineixen els corpus i la disciplina a què van lligada, la lingüística de corpus, és el fet de tenir un suport electrònic d'alguna mena. Per tant, alguns autors (Llamazares, 2008) consideren que la història dels corpus va començar a finals del segle XX, quan la lingüística estructural americana va establir les bases de la lingüística de corpus, mentre que d'altres (Francis, 1982) consideren que alguns reculls de textos de determinats autors o llengües es podria considerar un corpus i, com a resultat, la seva història començaria abans del segle XIX. Molt abans, de fet.

Segons Llamazares (2008) els corpus, abans del segle XIX, es definien per ser un conjunt de textos escrits amb la finalitat d'estudiar determinades llengües mortes, com ara el llatí. Francis (1992), un dels creadors d'un dels primers corpus electrònics en llengua anglesa, el BROWN, no va ser ell mateix qui va iniciar la història dels corpus. Francis (1982), va definir els corpus com un recull de textos que es representen una llengua o dialecte, o qualsevol altre subgrup d'una llengua per fer-lo servir per a anàlisis lingüístiques i s'acull a la seva pròpia definició per explicar que, ja al segle VIII hi havia reculls amb aquestes característiques. A més, McEnery i Hardie (2013) expliquen al seu llibre, *The History of Corpus Linguistics*, que hi va haver precursors a la lingüística de corpus que feien servir mètodes i reculls de textos similars als dels corpus molt abans. Va ser Roberto Busa (1951), qui va fer servir les primeres concordances generades per ordinador. A més, Alphonse Juilland (Juilland i Chang-Rodríguez, 1964) va establir els principis més importants a l'hora de treballar amb text electrònic, com ara la importància del mostreig equilibrat als corpus, o la importància de tenir en compte les estadístiques de dispersió i les freqüències brutes.

McEnery i Hardie (2013) expliquen com la lingüística de corpus no es va considerar tal fins a finals dels anys 50, a causa de la dependència gairebé absoluta que tenien els lingüistes en els textos informatitzats i els ordinadors. Fins que no es van desenvolupar aquestes tecnologies fins al punt de

poder gestionar i manipular una gran quantitat de text llegible per ordinador abans perquè es pogués crear el que actualment anomenem lingüística de corpus.

Entre els 60 i els 70 la lingüística de corpus va ser molt discontínua i pràcticament va desaparèixer a causa del racionalisme de Chomsky. L'aparició de Chomsky al panorama lingüístic i el seu racionalisme van provocar dures crítiques a la metodologia empírica dels corpus i durant gairebé dues dècades els corpus van ser desprestigiats. A més, altre lingüista expert i molt reconegut, Abercrombie (1965), va titllar l'ús dels corpus de "pseudotècniques", ja que considerava que el processament de dades era lent, amb molts errors i massa car per ser rendible. No obstant això, encara es van fer servir per a estudis de fonètica, adquisició de llengües i lingüística històrica, tot i amb l'arribada dels ordinadors la utilització de corpus va revifar (Llamazares, 2008).

El ressorgir d'aquesta lingüística de corpus es va donar als anys 80, quan diversos autors de renom, com G. Leech, van rebatre les crítiques teòriques i pràctiques d'Abercrombie i Chomsky. Els ordinadors es van convertir en un recurs indispensable per estudiar les llengües, fer hipòtesis lingüístiques i construir sistemes de processament del llenguatge natural.

Com bé expliquen Orăsan et al. (2007), els avanços en ordinadors i Internet van permetre que els investigadors poguessis processar corpus d'una gran mida sense necessitar un maquinari especialitzat, cosa que va ajudar als corpus a proliferar. A més, a finals dels 80 i principis dels 90 hi va haver un canvi en el paradigma i es va començar a fer recerques d'intel·ligència artificial, ja que van veure les limitacions dels sistemes d'aquell moment. En aquells anys van néixer els primers intents de motors estadístics i aprenentatge automàtic o *machine learning*.

McEnery i Hardie (2013) comenten que durant la primera dècada dels 2000 l'ús de corpus es va ampliar a altres àrees de coneixement i disciplines, com la lingüística contrastiva, l'anàlisi del discurs, l'aprenentatge de llengües, la semàntica, la sociolingüística, la lingüística teòrica, la lexicografia o la gramàtica, entre d'altres.

Durant les últimes dècades s'ha observat un augment en l'interès en la tecnologia lingüística basada en ordinadors, especialment en mètodes de reconeixement de llengües, reconeixement de parla (*speech recognition*), extracció de textos i informació automàtica, traducció automàtica, etc. Els corpus, actualment, s'utilitzen en àrees com, per exemple, el desenvolupament d'eines lingüístiques com ara bé correctors ortotipogràfics, sistemes de processament de paraules, sistemes d'edició i anàlisis de textos, entre d'altres. També es fan servir per desenvolupar diccionaris monolingües, bilingües i multilingües, tant impresos com digitals. En el món de la traducció són especialment útils,

també. Per exemple es fan servir per desenvolupar sistemes de suport a la traducció, sistemes d'accés a recursos lingüístics, sistemes de traducció automàtica, sistemes d'accés a informació multilingüe i sistemes de recuperació creuada. A més, s'estan desenvolupant noves tecnologies que necessiten gran quantitat de dades, com ara els sistemes d'OCR (*Optical Character Recognition*), sistemes de reconeixement de veu, sistemes de síntesi de la parla o *text-to-speech*, i sistemes d'aprenentatge web, entre d'altres (Dash i Arulmozi, 2018).

2.1.2. Classificació dels corpus

Hi ha una gran diversitat de característiques que diferencien els corpus entre si. Des del nombre d'elements que contenen, passant per quantes llengües hi consten a cada corpus o quin objectiu tenen. Als anys 90 i 2000 hi va haver un "boom" en la lingüística de corpus i la traducció automàtica i diversos lingüistes i pioners van crear les seves pròpies classificacions. Les que més destaquen són les d'EAGLES l'any 1996 i la de Torruella i Llisterri l'any 1999. Llamazares (2008) fa un recull d'aquestes classificacions i n'afegeix d'altres. Em basaré en l'article de Llamazares (2008), tot i que la classificació d'EAGLES (1996) va ser essencial en el seu moment. També ho va ser la de Torruella i Llisterri (1999), tres anys més tard. No obstant això, tant la classificació de Torruella i Llisterri com la d'EAGLES són dels anys noranta i l'article de Llamazares és una mica més actual i és més acurat pel que fa a l'estat de la tecnologia.

Segons Llamazares normalment els corpus es divideixen segons la modalitat de la llengua, segons el nombre de llegües, segons quin límit té un corpus, segons el caràcter, segons el període temporal, segons la mida i segons el tractament.

En primer lloc, parlarem de com es classifiquen els corpus segons la modalitat de la llengua. Principalment, hi ha tres tipus: escrits, orals i mixtos.

Corpus escrits: són corpus formats per textos escrits. N'és un exemple el CEDEL.

Corpus orals: són corpus que recullen mostres de la llengua parlada. N'és un exemple el COSER.

Corpus mixtos: són corpus que contenen tant mostres escrites com orals. Un bon exemple d'aquest tipus és el corpus CREA de la Real Academia Española.

A continuació, parlarem de com es classifiquen els corpus segons el nombre de llengües. La classificació té en compte el nombre de llengües i, en cas de ser més d'una, com es relacionen entre si.

Corpus monolingües: els corpus monolingües són corpus que contenen mostres d'una sola llengua o dialecte. Com per exemple el CREA o el CORGA.

Corpus bilingües: un corpus bilingüe és un corpus amb mostres de dues llengües que no necessàriament són una traducció l'una de l'altra.

Corpus multilingües: un corpus multilingüe és un corpus amb mostres de més de dues llengües que no necessàriament són traducció l'una de l'altra. Quan parlem d'un corpus amb més d'una llengua que compta amb segments originals i les seves traduccions se'n diuen **corpus paral·lels**. Els corpus paral·lels són molt importants per a la traducció automàtica, ja que amb ells s'entrenen els motors de traducció. I, a més, si aquestes traduccions estan alineades estem parlant de **corpus alineats**.
Pel que fa a la classificació de corpus segons els seus límits, podem observar tres tipus principals.

Corpus tancats: els corpus tancats són corpus que tenen un nombre finit de paraules, que s'acorda abans de fer el corpus. El corpus es dona per acabat quan s'ha assolit aquest nombre d'elements. N'és un exemple el BNC.

Corpus oberts: els corpus tancats són corpus que no tenen un nombre finit de paraules. Com, per exemple, el CREA o el Bank of English.

Corpus monitor: els corpus monitor són un tipus de corpus obert que destaca pel seu objectiu, que és actualitzar-se constantment. Originalment, els corpus monitor eren tancats, ja que se cercava tenir un corpus d'una mida constant perquè fos assequible per a les màquines d'aquell moment. Actualment, però, els corpus monitor són oberts, ja que la tecnologia pot suportar que un corpus es vagi actualitzant sense haver d'esborrar les dades antigues.

També es poden classificar segons el caràcter general o especialitzat dels textos

Corpus generals: els corpus generals tenen una gran varietat de gèneres textuals, modalitats, matèries, etc., ja que el que busquen és reflectir la llengua amb fidelitat.

Corpus especialitzats: els corpus especialitzats recullen textos que representin un aspecte concret d'una llengua, com ara un gènere textual. N'és exemple el Corpus Textual Especialitzat Plurilingüe de la Universitat Pompeu Fabra.

Una altra forma de classificar els corpus és tenint en compte el període temporal dels textos que els conformen.

Corpus diacrònics o històrics: els corpus diacrònics o històrics són corpus que inclouen textos de diferents èpoques amb l'objectiu d'observar l'evolució de la llengua. N'és un exemple el CORDE.

Corpus sincrònics: els corpus sincrònics són tot el contrari als diacrònics. Les mostres del corpus pertanyen totes a la mateixa època, ja que el que es pretén és observar diverses varietats lingüístiques en un moment determinat del temps.

Els corpus també es poden classificar tenint en compte si els textos que els conformen estan complets o no.

Corpus de referència: els corpus de referència són aquells que estan formats per fragments de textos per proporcionar informació molt completa sobre una llengua.

Corpus textuals: els corpus textuals inclouen textos sencers, sense fragmentar.

Finalment, els corpus es poden classificar segons el tractament que se li aplica al corpus, és a dir, si es deixa el corpus tal qual o se li afegeix informació metalingüística extra.

Corpus simples, no anotats o no codificats: els corpus no codificats són corpus amb text sense cap mena de format addicional, com ara codis o anotacions.

Corpus codificats o anotats: els corpus codificats estan formats per textos als quals se'ls hi ha afegit, manualment o automàticament, alguna mena d'informació extra. Poden ser etiquetes sobre l'autor, el títol, la categoria gramatical, etc. En són exemples la BDS (Base de Datos Sintácticos del Español Actual) o el CESS-ECE.

2.1.3. Criteris d'avaluació de corpus

Pel que fa a l'avaluació de la qualitat dels corpus, resulta sorprenent observar que no hi ha cap mena de mètriques, processos o mètodes exclusius de l'avaluació de corpus establerts en el sector que no depenguin, en certa manera, de la traducció automàtica. Hi ha una manca, potser no de recerca, però sí de recopilació i regulació.

Per avaluar els corpus paral·lels o bilingües, en la gran majoria de casos, com podem observar a diversos articles, s'apliquen mètodes d'avaluació de traducció automàtica, ja que l'objectiu final és el mateix: comparar com s'assemblen dos segments tenint en compte diferents factors.

Principalment, el que es fa és aplicar directament les mètriques d'avaluació de traducció automàtica al corpus. La mètrica que més s'utilitza amb diferència és el BLEU (Schwenk et al., 2021a; Zaragoza-Bernabeu et al., 2022; Laskar et al., 2020; Koehn et al., 2019). També observem en molts casos que els corpus s'empren per entrenar motors de TA (sobretot TAN, però alguns TAE) (Schwenk et al., 2021a; Zaragoza-Bernabeu et al., 2022; Laskar et al., 2020, Koehn et al., 2019; Defauw et al., 2019; Ramírez-Sánchez et al., 2020) i, posteriorment, avaluar amb mètriques els resultats d'aquests motors, a manera d'avaluació indirecta dels corpus.

A més a més, hi ha casos en què s'utilitzen programes com ara Bicleaner (Ramírez-Sánchez et al., 2020), Bicleaner AI (Zaragoza-Bernabeu et al., 2022), o motors i codificadors com RoBERTa (Artetxe et al., 2022), LASER i Moses (Chaudary et al., 2019), entre d'altres, o avaluació humana amb mètriques com la fluïdesa, la fidelitat o altres característiques, com les que ofereix l'eina KEOPS (Ramírez-Sánchez et al., 2020). No obstant això, aquestes opcions són una minoria comparades amb les dues mencionades anteriorment.

2.1.4. Eines per a neteja, avaluació o creació de corpus

2.1.4.1. Recollida de dades

- **Bitextor**

Bitextor[1] és una eina cofinançada per la Unió Europea que es fa servir per recollir bitexts de forma automàtica de pàgines web multilingüe.

- **Spiderling**

Spiderling[2] és altra eina que pertany a Corpus Tools. El seu objectiu és recollir parts de text enriquit d'internet i dades útils per als corpus de text.

- **wiki2corpus**

[1] https://github.com/bitextor/bitextor [Versió 8.3]
[2] https://corpus.tools/wiki/SpiderLing [Versió 2.2]

Wiki2corpus[3] és un script que descarrega articles de la Wikipedia de la llengua que se seleccioni i els extreu en forma prevertical que ajuda a processar les dades en altres eines de corpus. Funciona amb llicència.

2.1.4.2. Neteja i millora de corpus

- **MTUOC**

El projecte MTUOC[4] és un projecte creat per la UOC, en què s'inclou una sèrie de components que permeten entrenar i fer servir sistemes de traducció automàtica. Principalment consta de programes, mòduls i scripts de Python. El que resulta interessant per a aquest TFM és la quantitat d'elements de neteja i preprocessament de corpus que té.

El programa de neteja de corpus paral·lels, anomenat "MTUOC_clean_parallel_corpus.py", pot dur a terme gran quantitat d'accions, entre d'altres:

- ➔ Canviar l'apòstrof tipogràfic per l'estàndard
- ➔ Eliminar les etiquetes HTML i XMLE
- ➔ Eliminar els segments paral·lels en què alguna de les llengües tingui un segment buit
- ➔ Verificar les llengües de partida i d'arribada
- ➔ Aplicar expressions regulars

Pel que fa al preprocessament de corpus, MTUOC proporciona:

- ➔ Tokenitzadors en diverses llengües
- ➔ Un model de *truecasing*
- ➔ El programa de neteja de corpus MTUOC_clean_parallel_corpus.py
- ➔ Tractament d'expressions numèriques
- ➔ Càlcul de subparaules o *subwords*

- **Bifixer**

  Bifixer[5] és una eina de codi obert de neteja de bitexts cofinançada per la Unió Europea i que també forma part de les eines creades pels creadors de Bitextor. Entre les funcions de neteja de bitexts que té es poden destacar:

- ➔ Substituir caràcters d'alfabets incorrectes amb els correctes

[3] https://corpus.tools/wiki/wiki2corpus [Versió 2.0]
[4] https://xwiki.recursos.uoc.edu/wiki/mat00001ca/view/MTUOC%3A%20traducción%20automática%20neuronal/MTUOC%20-%20Programas%20y%20módulos%20de%20Python/ [Accedit el 12 de juliol de 2023]
[5] https://github.com/bitextor/bifixer [Versió 0.8.9]

- ➔ Normalitzar la puntuació i els espais
- ➔ Eliminar les etiquetes HTML
- ➔ Corregir errors comuns ortotipogràfics d'algunes llengües (com ara l'anglès, l'espanyol o l'alemany, entre d'altres)
- ➔ Eliminar els segments amb el *source* o el *target* buits.
- ➔ Corregir errors de tokenització
- ➔ Millorar la segmentació de segments amb frases llargues
- ➔ Netejar textos monolingües
- ➔ Marcar els segments duplicats
- ➔ etc.

- **BitextEdit**

BitextEdit és un model d'edició de bitexts que comenten Briakou et al. (2022b) per a la millora de la seva qualitat de forma automàtica. El que fa és agafar com a *input* un bitext i editar un dels dos segments per crear-ne una versió refinada. Es poden copiar bitexts de bona qualitat, editar parcialment errors de coincidència petits o traduir de zero referències incorrectes.

- **Zipporah**

Zipporah[6] és un sistema de neteja de corpus ideat concretament per a motors de traducció estadística. Aquesta eina se centra especialment en la selecció de dades "netes" d'un corpus amb molt de soroll amb l'objectiu d'utilitzar menys dades però de major qualitat a la llarga.

El funcionament d'aquesta eina és que, en primer lloc, es *map* o "mapegen" tots els parells de la proposta. Tot seguit s'entrena un model senzill de regressió per separar les dades bones de les dolentes o sintètiques. Una vegada entrenat el model, es fa servir per puntuar els parells de conjunts de dades. Finalment, els parells amb millor puntuació s'afegeixen a un subconjunt fins que s'arriba a la mida desitjada.

Com la fidelitat i la fluïdesa són els dos principals elements o característiques a l'hora de constituir un parell de bona qualitat, aquesta eina té dues funcions diferenciades per a cadascuna d'elles. Per a fidelitat es fa servir un sistema de puntuació *bag-of-words* i per a fluïdesa es fan servir puntuacions d'*n*-grama lingüístic.

[6] https://github.com/hainan-xv/zipporah [Accedit el 12 de juliol del 2023]

2.1.4.3. Avaluació de corpus

- **Bicleaner i Bicleaner AI**

Bicleaner[7] és una eina feta pel mateix equip que Bitextor. És una eina a Python de codi obert que classifica i neteja corpus paral·lels, si tenen segments amb soroll. Indica quina possibilitat hi ha que dos segments d'un corpus paral·lel siguin traduccions mútues establint valors del 0 a l'1, sent 0 traduccions totalment diferents i 1 sent traduccions mútues. Els parells que es considera que són soroll s'avaluaran amb un 0. També ofereix la possibilitat d'entrenar l'eina amb els paràmetres que vulguem per netejar els corpus.

Bicleaner AI[8] és l'última versió del Bicleaner, creada l'any 2022 diferent del Bicleaner tradicional. Aquesta actualització incorpora un classificador neuronal, amb l'objectiu de fer servir l'eina específicament per entrenar motors de traducció neuronal. Els experiments realitzats (Zaragoza-Bernabeu et al., 2022) demostren que els resultats milloren substancialment amb aquesta versió.

- **KEOPS**

El KEOPS[9] o *Keen Evaluator Of Parallel Sentences* és una eina de codi obert que permet fer una avaluació manual de frases paral·leles tant localment com amb *docker*. Permet fer diverses tasques, com ara gestionar projectes d'avaluació de corpus, anotar els segments, avaluar corpus seguint els criteris de fidelitat i fluïdesa de l'ELRC, tractar els resultats de l'avaluació, etc. Resulta especialment útil, ja que permet descarregar els resultats en format TSV.

2.1.4.4. Altres

- **TranslationDirection**

TranslationDirection[10] és una eina en procés de construcció. L'objectiu d'aquesta eina és entrenar un sistema que pugui distingir entre textos originals (OT), traduccions humanes (HT) i traduccions automàtiques (MT). Actualment, com està encara en procés de desenvolupament hi ha algunes funciones que no estan disponibles, però podria resultar molt útil per a la classificació de dades.

[7] https://github.com/bitextor/bicleaner [Versió 0.17.2]
[8] https://github.com/bitextor/bicleaner-ai [Versió 2.3.0]
[9] https://github.com/paracrawl/keops [Accedit el 12 de juliol de 2023]
[10] https://github.com/RikVN/TranslationDirection [Accedit el 12 de juliol de 2023]

- **Unitok**

Unitok[11] és un tokenitzador universal de text amb configuracions específiques per a gran quantitat de llengües. Pot convertir qualsevol text pla en una seqüència de *tokens* sense perdre les etiquetes tipus XML. Forma part d'una *suite* d'eines anomenada Corpus Tools.

- **Chared**

Chared[12] és una eina creada per detectar la codificació d'un text en una llengua coneguda. Té models per a gran varietat de llengües. També forma part de Corpus Tools.

- **JusText**

JusText[13] és una eina, part de Corpus Tools, d'eliminació de plantilles d'HTML que pot eliminar enllaços de navegació, capçaleres, peus de pàgina, etc. D'aquesta manera es deixa una pàgina de text simple amb frases completes.

- **Language Filter**

Language Filter[14] és una eina de discriminació lingüística que també forma part de Corpus Tools. Determina la llengua de paràgrafs i documents tenint en compte llistes predefinides de paraules amb freqüència de corpus.

- **Sketch Engine**

Sketch Engine[15] és una eina pionera i essencial del món dels corpus. La seva pàgina web proporciona corpus ja preparats per fer-los servir i eines per construir, actualitzar i instal·lar corpus. Entre les funcions que ens ofereix trobem:

  - ➔ *Word sketch:* funció que proporciona un resum del comportament gramatical i col·locacional d'una paraula
  - ➔ Concordança de paraules
  - ➔ Tesaurus
  - ➔ Identificació automàtica de llengües de corpus
  - ➔ Extracció de termes monolingüe

[11] https://corpus.tools/wiki/Unitok [Versió 3.2.7]
[12] https://corpus.tools/wiki/Chared [Versió 2.0]
[13] https://corpus.tools/wiki/Justext [Versió 3.0]
[14] https://corpus.tools/wiki/languagefilter [Versió 1.0]
[15] https://www.sketchengine.eu/#blue [Accedit el 12 de juliol de 2023]

➔ Extracció de termes bilingüe

➔ etc.

- **NoSketchEngine**

NoSketchEngine[16] és, com indica el nom, una versió més reduïda de l'Sketch Engine que han desenvolupat els creadors de Corpus Tools.

- **Biroamer**

El Biroamer[17] és una petita unitat que ajuda a anonimitzar o fer el que en anglès s'anomena ROAM (*Random, Omit, Anonymize and Mix*) els corpus paral·lels. És a dir, que l'eina agafa les dades que introduïm, les aleatoritza o "randomitza" i n'omet aproximadament un 10% i les barreja amb altre corpus. També és una eina creada per l'equip de Bitextor.

2.1.5. WikiMatrix

El corpus que s'utilitzarà en aquest TFM és WikiMatrix. Schwenk et al. (2021a) descriuen al seu article que WikiMatrix és un corpus multilingüe d'accés gratuït format per articles de la Wikipedia en un total de 85 llengües. No només s'inclouen les llengües més parlades, sinó que hi ha diversos dialectes o llengües amb pocs recursos. En total, quan el van fer l'any 2019, van aconseguir extreure 135 milions de segments paral·lels per a 1.620 parells de llengües diferents. Dels 135 milions de frases que van extreure, només 34 milions estaven alineades amb l'anglès.

L'extracció de les dades per a aquest corpus es va fer automàticament amb incrustació de paraules o *word embedding,* cosa que va permetre minar aquesta quantitat de dades sense tardar mesos i els resultats van ser tan positius que els motors de traducció que es van entrenar amb aquest corpus van aconseguir unes puntuacions de les mètriques BLEU bastant altes.

Segons els autors, van triar Wikipedia com a font dels segments perquè és un recurs gratuït d'internet amb un contingut molt divers i amb una gran diversitat de temes en més de 300 idiomes.

[16] https://nlp.fi.muni.cz/trac/noske [Accedit el 12 de juliol de 2023]
[17] https://github.com/bitextor/biroamer [Versió 2.10]

El procés que van seguir per preparar el text va ser el següent: en primer lloc, extreure el contingut textual. En segon lloc, dividir els paràgrafs en frases. En tercer lloc, eliminar les frases duplicades. I, per acabar, identificar les llengües de cada segment i eliminar les frases que no estan en la llengua en què haurien d'estar. Per a aquest projecte farem servir només el parell de llengües català-anglès.

2.2. Traducció automàtica

Per definir la traducció automàtica ens referirem al *Manual d'informàtica i de tecnologies per a la traducció* (2021) de Forcada. Forcada la defineix de la següent forma:

> *La traducció automàtica (TA) es pot definir com el procés (o el producte) de traduir un text informatitzat en una llengua origen a un text informatitzat en una llengua meta mitjançant l'ús d'un programa d'ordinador. Normalment, es reserva la denominació traducció automàtica per a la completament automàtica; quan s'hi produeix intervenció humana es parla de traducció assistida per l'ordinador o de traducció semi-automàtica. (Forcada, 2021, p.95)*

2.2.1. Tipus de traducció automàtica

Tot i que hi ha diferents classificacions de tipus de traducció automàtica em centraré en els tres principals: la TA basada en regles, la TA estadística i la TAN. Em basaré tant en Forcada (2021) com Oliver (2014) per definir aquests conceptes de manera concisa. A més, farem referència a un article web de Pestov (2018).

- **Traducció automàtica basada en regles (TABR)**

  En els sistemes de traducció automàtica basats en regles la informació necessària per a realitzar la traducció automàtica com, els diccionaris bilingües i les regles lingüístiques de cada llengua, les han escrit manualment experts lingüistes (Forcada, 2021). El seu punt àlgid va ser al principi de la creació de la traducció automàtica, entre 1950 i 1990, aproximadament. El seu ús s'ha vist bastant reduït, però encara s'utilitzen per a llengües minoritzades o amb poc recursos computacionals, com ara el català o l'occità. Com a exemple tenim Apertium.

- **Traducció automàtica estadística (TAE)**

  La traducció automàtica estadística entraria dins de la traducció automàtica basada en corpus. La TA basada en corpus obtén la informació que necessita per realitzar la traducció de manera inductiva a través de corpus paral·lels bilingües, que prèviament se segmenten i alineen (Forcada, 2021).

Pel que fa a la traducció automàtica estadística en concret, el que fa el motor és generar hipòtesis de traducció gràcies a càlculs estadístics i escull la traducció hipotètica més probable, la que tingui la puntuació més alta (Oliver, 2014). La traducció estadística va destacar entre els 90 i la dècada de 2010, tot i que actualment encara hi ha alguns motors que encara funcionen amb estadística (Pestov, 2018).

- **Traducció automàtica neuronal (TAN)**

  La traducció automàtica neuronal és una modalitat de traducció basada en l'anomenat aprenentatge profund o *deep learning,* on s'utilitzen mètodes d'un camp de la intel·ligència artificial anomenat *xarxes neurals* (Forcada, 2021). Dos exemples de traducció automàtica neuronal són Google Translator, Deepl o Yandex. És la modalitat que més es fa servir actualment. Des que van començar a sortir els primers motors TAN, entre l'any 2015 i 2016, ha anat pujant la seva popularitat cada vegada més, fins a arribar a ser la majoria dels motors de traducció utilitzats actualment (Pestov, 2018).

  Pel que fa a la relació entre la qualitat dels bitexts o corpus utilitzats i l'entrenament de motors de traducció automàtica neuronal, Khayrallah i Koehn (2018) demostren l'efecte que tenen els diferents tipus de soroll als motors TAN, com ara la còpia, les llengües errònies, el contingut no lingüístic, els segments buits, etc. A més, Lample et al. (2018) afegeixen que la incertesa de les dades que venen de referències on hi ha soroll, contribueixen de gran manera en el calibratge incorrecte dels models TAN. Els errors de significat petits fan que els motors es degenerin i no tinguin unes prediccions tan segures, afirmen Briakou et al. (2021).

  A l'hora de tractar amb aquests errors i millorar els corpus, hi ha moltes estratègies diferents, com ara fer servir regles de filtratge predefinides basades en identificadors de llengües (Rossenbach et al., 2018), tenir en compte les *scoring functions* basades en models lingüístics, extreure funcions dels models TAN i les probabilitats lèxiques de la traducció (Sánchez-Cartagena et al., 2018), combinar incrustacions de paraules o *word embeddings* ja entrenades prèviament (Papavassiliou et al., 2018), entropia dual (Chaudhary et al., 2019) i revisar les traduccions imperfectes dels bitexts amb traduccions sintètiques generades per TAN de bona qualitat (Briakou et al., 2022a), entre d'altres.

Per a aquest projecte el que més ens interessa és la traducció automàtica neuronal, ja que serà on, amb sort, acabarà el corpus funcional de WikiMatrix.

2.2.2. La traducció automàtica per a llengües minoritzades i la neteja de corpus

A més, la traducció automàtica per a llengües minoritzades és un repte, ja que hi ha molt pocs bitexts, és a dir, textos traduïts en dos idiomes (Koehn i Knowles, 2017). Els models de traducció automàtica normalment s'entrenen alineant-los heurísticament (Esplà et al., 2019) o amb dades minades automàticament (Schwenk et al., 2021a,b), cosa que fa que els resultats siguin de mala qualitat, ja que poden incloure diferències de significat, traduccions incorrectes, llengües externes al parell del corpus, caràcters erronis, soroll, segments buits, etc.

La neteja i filtració de corpus extrets d'internet és una pràctica que s'ha estandarditzat força a l'hora de crear corpus d'alta qualitat per poder entrenar els models de traducció automàtica i que donin resultats correctes (Koehn et al. 2018). Resulta molt útil a l'hora de tractar amb pocs recursos (Koehn 2020), com en el cas del català.

No obstant això, hi ha experts en el sector (Briakou et al., 2022a) que consideren que té dues limitacions principals. Exposen que si eliminem les traduccions que només són parcialment correctes, per evitar que "embrutin" el motor, i filtrem segments, perdem molta informació i contingut que és essencial en el cas de les llengües minoritzades. Proposen l'ús d'eines d'edició automàtica de bitexts, com ara BitextEdit.

Els bitexts, o corpus, extrets amb mineria com ara els de CCAligned (Koehn et al. 2020), WikiMatrix (Schwenk et al., 2021a) i ParaCrawl (Zaragoza-Bernabeu et al., 2020; Esplà et al. 2019) tenen errors sistemàtics, especialment amb llengües minoritzades i la gran majoria de parells amb pocs recursos tenen menys d'un 50% de traduccions vàlides (Briakou et al., 2020). Fins i tot amb parells de molts recursos, com ara l'anglès-francès, podem trobar fins a un 40% d'errors en aquest tipus de corpus.

2.2.3. Mètriques d'avaluació de traducció automàtica

Com bé indiquen Sánchez i Rico (2020), hi ha dos tipus d'avaluació de TA: l'avaluació manual, realitzada per lingüistes, i l'automàtica, realitzada per un programa informàtic sense intervenció humana.

Quan parlem d'avaluació manual o humana, ens referim al fet que uns lingüistes mesuren la qualitat principalment en termes de fidelitat i fluïdesa (*adequacy* i *fluency*) (Sánchez i Rico, 2020). Segons Moorkens et al. (2018), hi ha més factors avaluables a part de la fidelitat i la fluïdesa, com ara la precisió, la llegibilitat, la comprensió i l'acceptabilitat. Clarament, avaluar manualment la qualitat d'un

corpus o del resultat d'una traducció automàtica és llarg i costós, tot i que permet aprofundir més en els errors.

Quan parlem d'avaluació automàtica, ens referim a la comparació de dues traduccions, i la traducció que més s'apropi a la traducció de referència és la que es considerarà millor (Sánchez i Rico, 2020).

Lavie (2011) comenta alguns dels paràmetres que els programes tenen en compte a l'hora d'assignar una puntuació a cada parell:

- **Precisió**: ràtio entre paraules correctes de la traducció i el nombre total de paraules.
- **Exhaustivitat**: ràtio entre paraules correctes en la traducció i el nombre total de paraules en la traducció de referència.
- **Distància d'edició de Levenshtein**: nombre d'insercions, eliminacions i substitucions que necessita la traducció automàtica per a ser igual que la traducció de referència.

A continuació s'explicaran resumidament les mètriques de traducció automàtica més populars:

BLEU (*Bilingual Evaluation Understudy*): és la més popular actualment. BLEU (Papieni et al., 2002) té en compte els següents factors per computar la seva puntuació, que va de 0 a 1 (essent el 0 la qualitat més baixa amb una traducció totalment diferent de l'original i 1 la qualitat més alta amb una traducció idèntica a l'original):

- Que l'extensió de la traducció sigui igual a l'original o referència
- Que les paraules de la traducció siguin iguals a les de l'original o referència
- Que l'ordre de les paraules sigui igual que en l'original o referència

METEOR (*Metric For Evaluation of Translation with Explicit Ordering*): Banerjee i Lavie presenten l'any 2005 una mètrica que, en lloc de calcular els *n*-grames, es basa en el que els seus autors anomenen unigrames. Segons He i Way (2009) METEOR valora millor les traduccions més llargues.

TER (*Translation Edit Rate* o *Translation Error Rate*): TER (Snover et al., 2006) és una mètrica que mesura el percentatge de caràcters que s'hauran de modificar perquè la traducció es correspongui amb la traducció de referència. Segons He i Way (2009) TER prefereix traduccions més curtes.

NIST (*National Institute of Standards and Technology*): Doddington (2002) va crear NIST com a una modificació de BLEU que li dona un pes especial a les paraules poc emprades.

WER (*Word Error Rate*): WER és una variació de TER creada per Ye-Yi Wang et al. (2003) que mesura a nivell de paraula en lloc de caràcters. Té en compte les insercions, eliminacions i substitucions de paraules, així com l'ordre de les paraules.

chrF3 (*character n-gram F-score*): és una mètrica d'avaluació alternativa a BLEU que es calcula segons el valor-f de l'*n*-grama a nivell de caràcters. És independent a la llengua i no necessita una tokenització.

BEER (*BEtter Evaluation as Ranking*): segons Stanojević i Sima'an (2015), BEER és una mètrica que combina tant els n-grames com els arbres de permutacions.

3. Metodologia

3.1. Introducció

Com ja sabia que el corpus de WikiMatrix contenia molts errors i la puntuació era dolenta, l'objectiu era esbrinar, mitjançant l'avaluació del corpus, si Bicleaner era una eina efectiva per a netejar i filtrar corpus. Per fer això, vaig avaluar el corpus amb dos programes diferents: Bicleaner, que avaluava el corpus de forma automàtica, i KEOPS, amb què es podien puntuar els segments de forma manual.

El corpus de WikiMatrix anglès-català que vaig utilitzar prové d'OPUS[18], un repositori en línia gratuït de corpus paral·lels extrets d'Internet i amb arxius ".tsv.gz." Això volia dir que, una vegada descomprimit, el corpus quedaria en un format de text separat per tabuladors, també conegut com TSV. El corpus constava d'una primera columna amb la puntuació de la qualitat, una segona columna amb el segment català i una tercera columna amb el segment en anglès, separades per tabuladors.

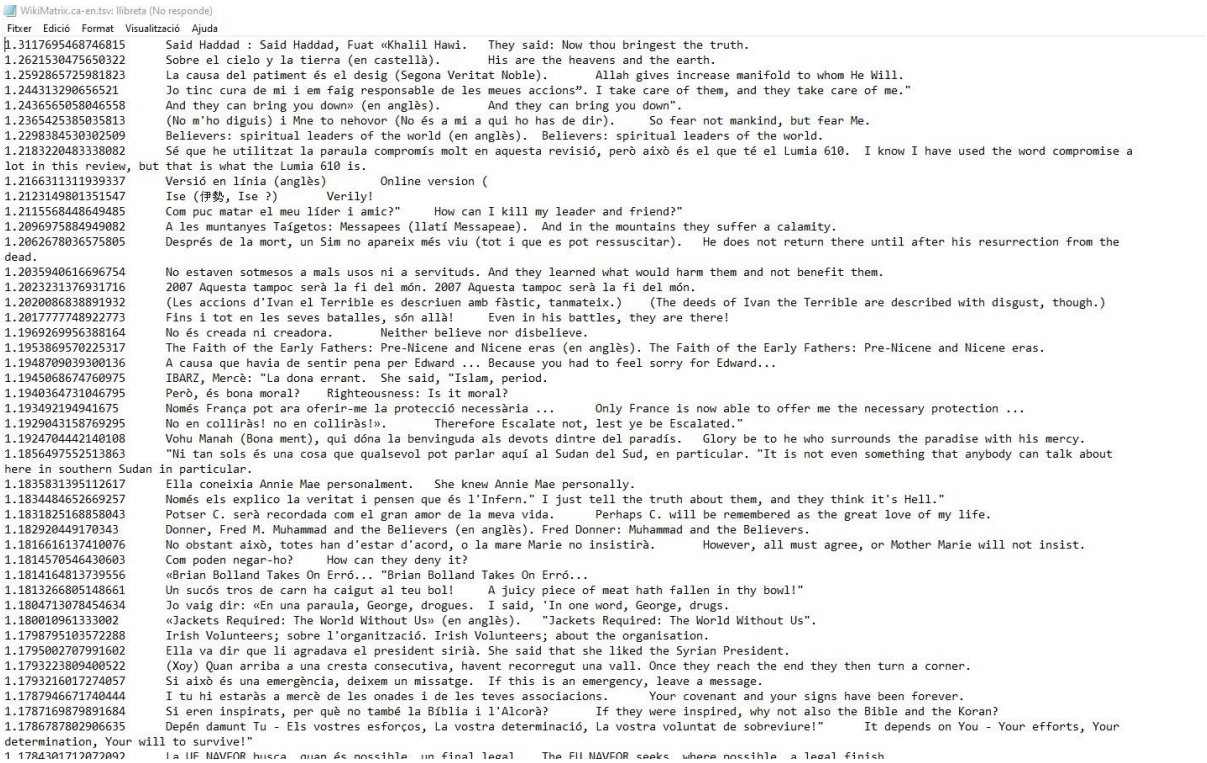

WikiMatrix.ca-en.tsv: llibreta (No responde)
Fitxer Edició Format Visualització Ajuda
1.3117695468746815 Said Haddad : Said Haddad, Fuat «Khalil Hawi. They said: Now thou bringest the truth.
1.2621530475650322 Sobre el cielo y la tierra (en castellà). His are the heavens and the earth.
1.2592865725981823 La causa del patiment és el desig (Segona Veritat Noble). Allah gives increase manifold to whom He Will.
1.244313290656521 Jo tinc cura de mi i em faig responsable de les meues accions". I take care of them, and they take care of me."
1.2436565058046558 And they can bring you down» (en anglès). And they can bring you down".
1.2365425385035813 (No m'ho diguis) i Mne to nehovor (No és a mi a qui ho has de dir). So fear not mankind, but fear Me.
1.2298384530302509 Believers: spiritual leaders of the world (en anglès). Believers: spiritual leaders of the world.
1.2183220483338082 Sé que he utilitzat la paraula compromís molt en aquesta revisió, però això és el que té el Lumia 610. I know I have used the word compromise a lot in this review, but that is what the Lumia 610 is.
1.2166311311939337 Versió en línia (anglès) Online version (
1.2123149801351547 Ise (伊勢, Ise ?) Verily!
1.2115568448649485 Com puc matar el meu líder i amic?" How can I kill my leader and friend?"
1.2096975884949082 A les muntanyes Taígetos: Messapees (llatí Messapeae). And in the mountains they suffer a calamity.
1.2062678036575805 Després de la mort, un Sim no apareix més viu (tot i que es pot ressuscitar). He does not return there until after his resurrection from the dead.
1.2035940616696754 No estaven sotmesos a mals usos ni a servituds. And they learned what would harm them and not benefit them.
1.2023231376931716 2007 Aquesta tampoc serà la fi del món. 2007 Aquesta tampoc serà la fi del món.
1.2020086838891932 (Les accions d'Ivan el Terrible es descriuen amb fàstic, tanmateix.) (The deeds of Ivan the Terrible are described with disgust, though.)
1.2017777748922773 Fins i tot en les seves batalles, són allà! Even in his battles, they are there!
1.1969269956388164 No és creada ni creadora. Neither believe nor disbelieve.
1.1953869570225317 The Faith of the Early Fathers: Pre-Nicene and Nicene eras (en anglès). The Faith of the Early Fathers: Pre-Nicene and Nicene eras.
1.1948709039300136 A causa que havia de sentir pena per Edward ... Because you had to feel sorry for Edward...
1.19450686747760975 IBARZ, Mercè: "La dona errant. She said, "Islam, period.
1.1940364731046795 Però, és bona moral? Righteousness: Is it moral?
1.193492194941675 Només França pot ara oferir-me la protecció necessària ... Only France is now able to offer me the necessary protection ...
1.1929043158769295 No en colliràs! no en colliràs!». Therefore Escalate not, lest ye be Escalated."
1.1924704442140108 Vohu Manah (Bona ment), qui dóna la benvinguda als devots dintre del paradís. Glory be to he who surrounds the paradise with his mercy.
1.1856497552513863 "Ni tan sols és una cosa que qualsevol pot parlar aquí al Sudan del Sud, en particular. "It is not even something that anybody can talk about here in southern Sudan in particular.
1.1835831395112617 Ella coneixia Annie Mae personalment. She knew Annie Mae personally.
1.1834484652669257 Només els explico la veritat i pensen que és l'Infern." I just tell the truth about them, and they think it's Hell."
1.1831825168858043 Potser C. serà recordada com el gran amor de la meva vida. Perhaps C. will be remembered as the great love of my life.
1.182920449170343 Donner, Fred M. Muhammad and the Believers (en anglès). Fred Donner: Muhammad and the Believers.
1.18166616137410076 No obstant això, totes han d'estar d'acord, o la mare Marie no insistirà. However, all must agree, or Mother Marie will not insist.
1.1814570546430603 Com poden negar-ho? How can they deny it?
1.18141648413739556 «Brian Bolland Takes On Erró... "Brian Bolland Takes On Erró...
1.1813266805148661 Un sucós tros de carn ha caigut al teu bol! A juicy piece of meat hath fallen in thy bowl!"
1.1804713078454634 Jo vaig dir: «En una paraula, George, drogues. I said, 'In one word, George, drugs.
1.180010961333002 «Jackets Required: The World Without Us» (en anglès). "Jackets Required: The World Without Us".
1.1798795103572288 Irish Volunteers; sobre l'organització. Irish Volunteers; about the organisation.
1.1795002707991602 Ella va dir que li agradava el president sirià. She said that she liked the Syrian President.
1.1793223809400522 (Xoy) Quan arriba a una cresta consecutiva, havent recorregut una vall. Once they reach the end they then turn a corner.
1.1793216017274057 Si això és una emergència, deixem un missatge. If this is an emergency, leave a message.
1.1787946671740444 I tu hi estaràs a mercè de les onades i de les teves associacions. Your covenant and your signs have been forever.
1.1787169879891684 Si eren inspirats, per què no també la Bíblia i l'Alcorà? If they were inspired, why not also the Bible and the Koran?
1.1786787802906635 Depén damunt Tu - Els vostres esforços, La vostra determinació, La vostra voluntat de sobreviure!" It depends on You - Your efforts, Your determination, Your will to survive!"
1.1784301712072092 La UE NAVFOR busca quan és possible un final legal The EU NAVFOR seeks where possible a legal finish

Figura 1. Corpus WikiMatrix

[18] https://opus.nlpl.eu [Accedit 12 de juliol de 2023]

3.2. Preparació del corpus

Abans de començar a instal·lar i fer servir el Bicleaner, el primer que s'havia de fer era convertir el corpus a un arxiu de text de 4 columnes:

1. URL 1
2. URL 2
3. Segment d'origen
4. Segment de destí

Al GitHub del Bicleaner s'indicava que així havia de ser l'arxiu *input* perquè el programa el pogués processar. Ho vaig fer amb ajuda del llenguatge AWK de Linux, que està dissenyat per processar dades basades en text. Aquest llenguatge tenia moltíssimes opcions per a textos amb columnes. No obstant això, com no estava gens familiaritzada amb aquest llenguatge vaig haver de fer diversos intents fins que vaig aconseguir convertir el corpus de 3 columnes (la puntuació de qualitat automàtica original, el català i l'anglès) en un document de 4 columnes, que vaig decidir que fossin:

1. El nom del corpus: WikiMatrix
2. La puntuació original de qualitat
3. El segment en català
4. El segment en anglès

El comandament que vaig introduir va ser el següent:

```
head -10 WikiMatrix.ca-en.tsv > testsample.txt
awk '{print $1, $2, $3}' testsample.txt
```

Primer vaig fer servir el comandament "head -n", on "head" fa que s'extreguin els elements del principi del document i "n" es substitueix pel nombre d'elements que es volen extreure. Tot seguit vaig introduir el nom del fitxer original "WikiMatrix.ca-en.tsv" i el nom del fitxer final que es crearia "testsample.txt", amb el símbol ">" pel mig, que serveix per indicar on es volen posar les dades. Amb aquest comandament volia extreure 10 línies per fer un experiment per trobar quin comandament era el que necessitava per modificar tot el corpus i desar el resultat en un arxiu de text diferent. Després vaig posar un comandament d'AWK on "print" fa que el que s'indica a continuació es mostri en pantalla i els números precedits del símbol "$" representen les columnes que volem mostrar. El resultat havia de ser que es mostressin en pantalla la primera, la segona i la tercera columna del fitxer que havia creat amb el comandament anterior.

No obstant això, el primer intent no va donar els resultats esperats, ja que la segona columna es tallava per la meitat.

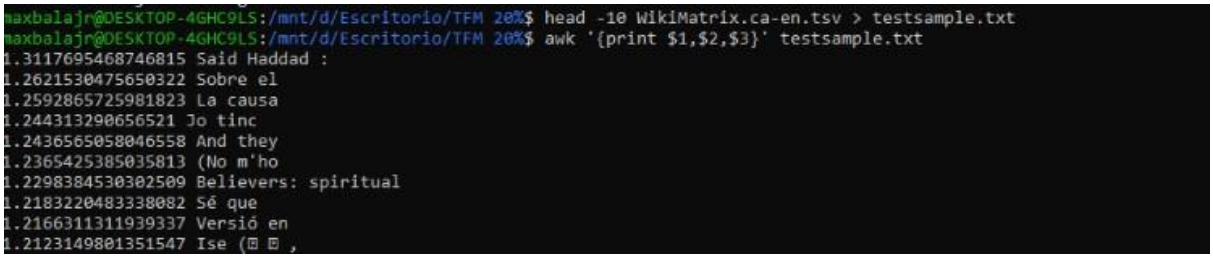


Figura 2. Primer intent de modificar el corpus amb AWK

En el següent intent vaig afegir un comandament, que teòricament li indicava a Linux que havia de separar les columnes per tabuladors amb "-F" i, a més, afegia la columna una columna davant de la columna "1" amb el nom del corpus. Per crear una columna que no existeix al fitxer original que es manipula amb una cadena o *string* de text determinada s'ha d'afegir la cadena de text entre cometes al lloc on es vol que aparegui. El comandament resultant va ser aquest:

```
awk -F '\t' '{print "WikiMatrix" $1, $2, $3}' testsample.txt
```

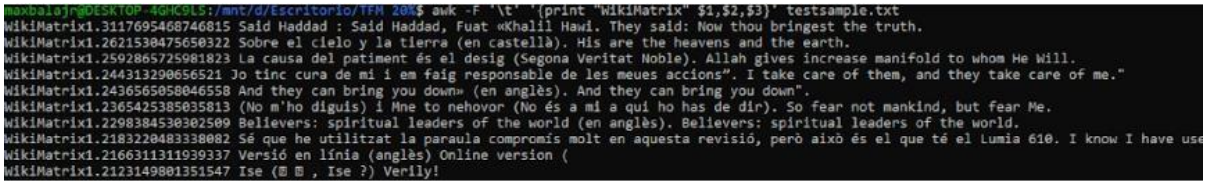


Figura 3. Segon intent de modificar el corpus amb AWK

Així i tot, el resultat no va ser l'esperat tampoc, ja que, tot i que aquesta vegada mostrava tot el text de totes les columnes i tenia la columna de WikiMatrix, continuava sense separar-lo per tabuladors. Finalment, vaig provar una variació d'aquest mateix comandament que sí que va funcionar:

```
awk -F "\t"{FS=OFS="\t"} {print "WikiMatrix" $1, $2, $3 > (baeldung.sample.txt)}' WikiMatrix.ca-en.tsv
```

El que fa "\t" és indicar que el separador que es vol utilitzar és el tabulador, "FS" llegeix i agafa les variables que hem indicat prèviament amb el símbol "$" i hi aplica el separador, i "OFS" indica que això ha de ser l'*output.* Per fer servir aquestes variables noves vaig haver de canviar l'ordre i la manera

en què indicava quin era el fitxer original, que vaig haver de posar al final del comandament, i quin havia de ser el resultat, que vaig haver d'indicar amb uns parèntesis.

Aquest comandament sí que va funcionar, pel que el vaig poder aplicar a tot el corpus. D'aquesta manera ja tenia el corpus amb el nombre de columnes que calia per introduir-lo al Bicleaner.

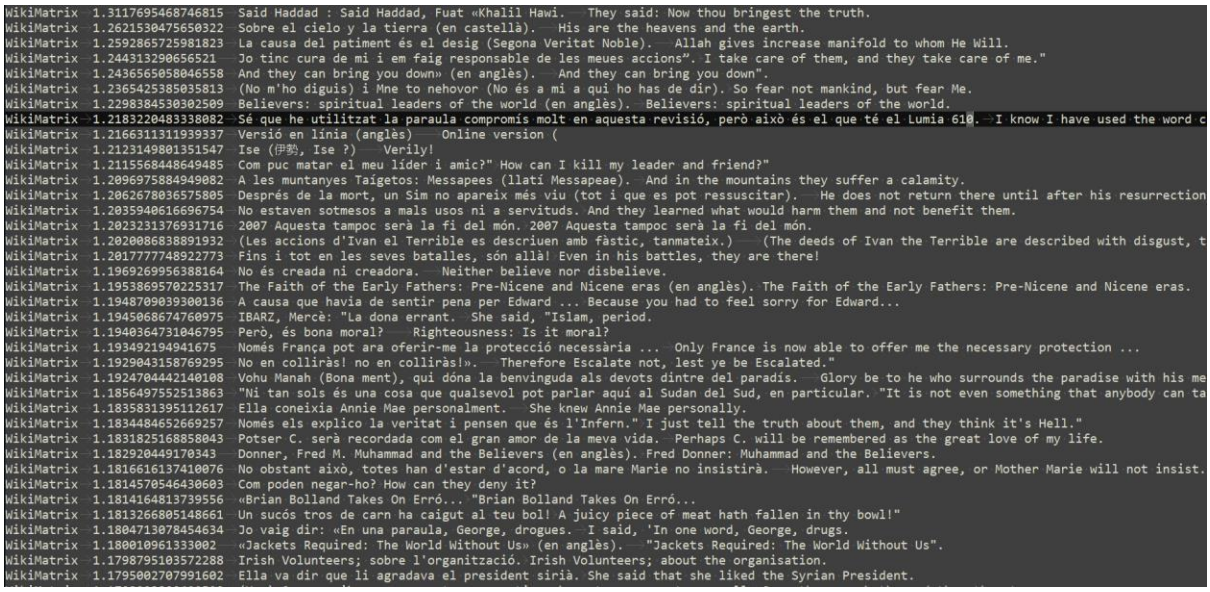


Figura 4. Corpus final modificat amb AWK

3.3. Avaluació automàtica amb Bicleaner

Com he explicat anteriorment, Bicleaner és una eina a Python que classifica i neteja corpus paral·lels, si tenen segments amb soroll. El que fa és indicar quina possibilitat hi ha que dos segments d'un corpus paral·lel siguin traduccions mútues establint valors del 0 a l'1, sent 0 traduccions totalment diferents i 1 sent iguals. També ofereix la possibilitat d'entrenar l'eina amb els paràmetres que vulguem per netejar els corpus, però jo vaig fer servir la configuració que hi havia per defecte.

Per poder instal·lar aquest programa calia fer-ho des de la terminal. En el meu cas vaig fer servir el Linux al meu ordinador amb Windows amb el comandament "pip", ja que, quan vaig intentar instal·lar-lo fent servir la consola CMD de Windows i el PowerShell donava error.

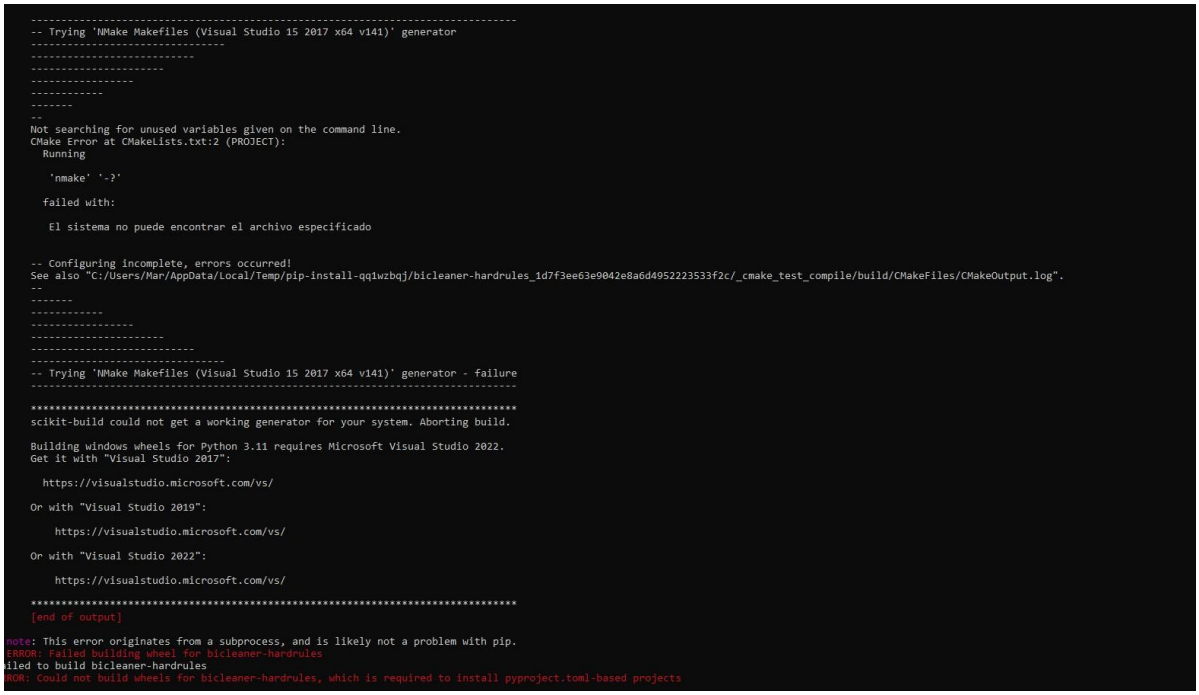


Figura 5. Captura d'un error a l'hora d'instal·lar el Bicleaner

Figura 6. Captura d'un error a l'hora d'instal·lar el Bicleaner 2

Per instal·lar-lo calia fer servir els comandaments següents, que s'indicaven a l'apartat "Installation & Requirements" del GitHub del Bicleaner:

```
pip install bicleaner
pip install --config-settings="--build-option=--max_order=7"
```

No obstant això, per poder instal·lar el Bicleaner amb Linux primerament vaig haver d'actualitzar Ubuntu amb ajuda del comandament:

```
sudo apt && sudo apt upgrade -y
```

Tot seguit vaig haver d'instal·lar Hunspell, que era necessari perquè un dels mòduls funcionés, però no s'indicava a les instruccions d'instal·lació del Bicleaner. Vaig fer servir el comandament següent:

```
sudo ap install hunspell libhunspell-dev
```

Finalment, vaig tornar a intentar el comandament "pip install bicleaner" amb què s'instal·lava el Bicleaner i els mòduls que necessitava per poder funcionar i amb el comandament "sudo ldconfig".

No obstant això, vaig tenir problemes per instal·lar el programa. Em donava l'error següent:

```
PermissionError: [Errno 13] Permission denied:
'/usr/local/lib/python3.8/dist-packages/fastspell/lid.176.bin'
```

```
maxbalajr@DESKTOP-4GHC9LS:/mnt/d/Escritorio/TFM 20%/Bicleaner$ bicleaner-classify sample.txt sampleclassified.txt en-ca.yaml
2023-03-14 20:17:35,781 - INFO - Accuracy histogram: [0.5000758, 0.7002882, 0.7630821, 0.8126801, 0.8516609, 0.8691036, 0.8471106, 0.7600485, 0.6203549, 0.5078113]
2023-03-14 20:17:35,781 - INFO - Ideal threshold: 0.5
2023-03-14 20:17:35,781 - WARNING - Porn removal not present in metadata, disabling.
2023-03-14 20:17:35,796 - INFO - Arguments processed.
2023-03-14 20:17:35,796 - WARNING - Downloading FastText model...
Traceback (most recent call last):
  File "/usr/local/bin/bicleaner-classify", line 8, in <module>
    sys.exit(main())
  File "/usr/local/lib/python3.8/dist-packages/bicleaner/bicleaner_classifier_full.py", line 246, in main
    args = initialization() # Parsing parameters
  File "/usr/local/lib/python3.8/dist-packages/bicleaner/bicleaner_classifier_full.py", line 49, in initialization
    fs_obj = FastSpell("en", mode="aggr")
  File "/usr/local/lib/python3.8/dist-packages/fastspell/fastspell.py", line 76, in __init__
    self.download_fasttext()
  File "/usr/local/lib/python3.8/dist-packages/fastspell/fastspell.py", line 91, in download_fasttext
    urllib.request.urlretrieve(self.ft_download_url, ft_model_path)
  File "/usr/lib/python3.8/urllib/request.py", line 257, in urlretrieve
    tfp = open(filename, 'wb')
PermissionError: [Errno 13] Permission denied: '/usr/local/lib/python3.8/dist-packages/fastspell/lid.176.bin'
```

Figura 7. Captura de l'error PermissionError a l'hora d'instal·lar el Bicleaner

A més a més, vaig haver d'instal·lar KenLM amb el comandament:

```
sudo pip install https://github.com/kpu/kenlm/archive/master.zip
```

Una vegada instal·lat vaig procedir a executar el Bicleaner amb "bicleaner-classify", però tot i així em donava "ERROR - Program finished with errors"

El Bicleaner esperava primer la columna d'anglès i després la columna del català, així que vaig haver de refer el corpus amb AWK per arreglar-ho, canviant l'ordre de les columnes "2" i "3". En aquesta ocasió també vaig fer una prova que vaig anomenar "columnesordenades.txt". Una vegada vaig comprovar que el resultat era satisfactori, vaig repetir el comandament, aquesta vegada amb el fitxer de sortida "corpusordenat.txt". El comandament que vaig utilitzar va ser el següent:

```
awk -F "\t"{FS=OFS="\t"} {print "WikiMatrix" $1, $3, $2 > (columnes ordenades.txt)}' WikiMatrix.ca-en.tsv
```

A més, hi havia diverses versions de Python instal·lades, així que vaig haver d'actualitzar el Python a 3.11 i reinstal·lar el Bicleaner amb els següents comandaments:

1. `sudo apt remove python`
2. `sudo add-apt-repository ppa:deadsnakes/ppa && sudo apt update`
3. `sudo apt install python3.11 python-pip`
4. `sudo python3.11 -m pip install --upgrade pip`

A més, vaig instal·lar els comandaments anteriors amb "sudo" davant, ja que si no hi posava aquest comandament donaven error. Finalment vaig aconseguir resoldre tots els problemes.

Vaig executar el següent comandament:

```
sudo bicleaner-classify columnesordenades.txt sampleclassified.txt en-ca.yaml
```

L'objectiu era comprovar si funcionava bé el programa amb una mostra de 50 línies, que ja havia extret prèviament amb el mateix comandament "head -n" que vaig anomenar "sampleclassified.txt". Una vegada comprovat que sí que funcionava, vaig executar el comandament amb el corpus sencer amb el comandament:

```
sudo bicleaner-classify corpusordenat.txt corpusclassificat.txt en-ca.yaml
```

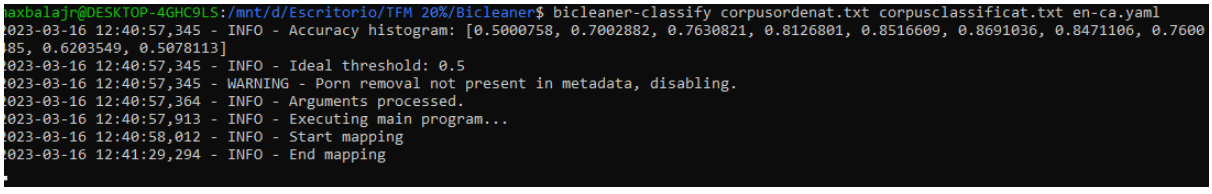


Figura 8. Aplicació del comandament bicleaner-classify al corpus

Una vegada aplicat el classificador del Bicleaner, el document final passava a tenir 5 columnes:

1. El nom del corpus: WikiMatrix
2. La puntuació original de qualitat
3. L'anglès

4. El català
5. La puntuació del Bicleaner

WikiMatrix	1.31176954687468615	They said: Now thou bringest the truth. Said Haddad : Said Haddad, Fuat «Khalil Hawi.	0
WikiMatrix	1.2621530475650322	His are the heavens and the earth.	Sobre el cielo y la tierra (en castellà).	0
WikiMatrix	1.2592865725981823	Allah gives increase manifold to whom He Will.	La causa del patiment és el desig (Segona Veritat Noble).	0.458
WikiMatrix	1.244313290656521	I take care of them, and they take care of me."	Jo tinc cura de mi i em faig responsable de les meues accions".	0.488
WikiMatrix	1.2436565058046558	And they can bring you down".	And they can bring you down» (en anglès).	0.694
WikiMatrix	1.2365425385035813	So fear not mankind, but fear Me.	(No m'ho diguis) i Mne to nehovor (No és a mi a qui ho has de dir).	0
WikiMatrix	1.2298384530302509	Believers: spiritual leaders of the world.	Believers: spiritual leaders of the world (en anglès).	0.822
WikiMatrix	1.2183220483338082	I know I have used the word compromise a lot in this review, but that is what the Lumia 610 is.	Sé que he utilitzat la paraula compromís molt en aquesta revisió, però això	
610.	0.820			
WikiMatrix	1.2166311311939337	Online version (	Versió en línia (anglès)	0
WikiMatrix	1.2123149801351547	Verily! Ise (伊勢, Ise ?)	0	
WikiMatrix	1.2115568448649485	How can I kill my leader and friend?"	Com puc matar el meu líder i amic?"	0.648
WikiMatrix	1.2096975884949082	And in the mountains they suffer a calamity.	A les muntanyes Taigetos: Messapees (llatí Messapeae).	0.594
WikiMatrix	1.2062678036575805	He does not return there until after his resurrection from the dead.	Després de la mort, un Sim no apareix més viu (tot i que es pot ressuscitar).	0.304
WikiMatrix	1.2035940616696754	And they learned what would harm them and not benefit them.	No estaven sotmesos a mals usos ni a servituds. 0.504	
WikiMatrix	1.202323137693171 6	2007 Aquesta tampoc serà la fi del món. 2007 Aquesta tampoc serà la fi del món. 0		
WikiMatrix	1.2020086838891932	(The deeds of Ivan the Terrible are described with disgust, though.)	(Les accions d'Ivan el Terrible es descriuen amb fàstic, tanmateix.)	0.918
WikiMatrix	1.2017777748922773	Even in his battles, they are there!	Fins i tot en les seves batalles, són allà!	0.680
WikiMatrix	1.1969269956388164	Neither believe nor disbelieve. No és creada ni creadora.	0.724	
WikiMatrix	1.1953869570225317	The Faith of the Early Fathers: Pre-Nicene and Nicene eras.	The Faith of the Early Fathers: Pre-Nicene and Nicene eras (en anglès). 0.718	
WikiMatrix	1.1948709039300136	Because you had to feel sorry for Edward...	A causa que havia de sentir pena per Edward ... 0.604	
WikiMatrix	1.1945068674760975	She said, "Islam, period.	IBARZ, Mercè: "La dona errant. 0.694	
WikiMatrix	1.1940364731046795	Righteousness: Is it moral?	Però, és bona moral? 0.708	
WikiMatrix	1.193492194941675	Only France is now able to offer me the necessary protection ...	Només França pot ara oferir-me la protecció necessària ...	0.730
WikiMatrix	1.1929043158769295	Therefore Escalate not, lest ye be Escalated."	No en colliràs! no en colliràs!».	0.462
WikiMatrix	1.1924704442140108	Glory be to he who surrounds the paradise with his mercy.	Vohu Manah (Bona ment), qui dóna la benvinguda als devots dintre del paradís.	0.238
WikiMatrix	1.1856497552513863	"It is not even something that anybody can talk about here in southern Sudan in particular.	"Ni tan sols és una cosa que qualsevol pot parlar aquí al Sudan del Sud, en	
WikiMatrix	1.1835831395112617	She knew Annie Mae personally.	Ella coneixia Annie Mae personalment. 0.780	
WikiMatrix	1.1834484652669257	I just tell the truth about them, and they think it's Hell."	Només els explico la veritat i pensen que és l'Infern." 0.734	
WikiMatrix	1.1831825168858043	Perhaps C. will be remembered as the great love of my life.	Potser C. serà recordada com el gran amor de la meva vida.	0.648
WikiMatrix	1.182920449170343	Fred Donner: Muhammad and the Believers.	Donner, Fred M. Muhammad and the Believers (en anglès). 0.806	
WikiMatrix	1.1816616137410076	However, all must agree, or Mother Marie will not insist.	No obstant això, totes han d'estar d'acord, o la mare Marie no insistirà.	0.598
WikiMatrix	1.1814570546430603	How can they deny it?	Com poden negar-ho? 0.536	
WikiMatrix	1.1814164813739556	"Brian Bolland Takes On Erró... «Brian Bolland Takes On Erró... 0		
WikiMatrix	1.1813266805148661	A juicy piece of meat hath fallen in thy bowl!"	Un sucós tros de carn ha caigut al teu bol!	0.448
WikiMatrix	1.1804713078454634	I said, 'In one word, George, drugs.	Jo vaig dir: «En una paraula, George, drogues. 0.762	
WikiMatrix	1.180010961333002	"Jackets Required: The World Without Us".	«Jackets Required: The World Without Us» (en anglès).	0.752
WikiMatrix	1.1798795103572288	Irish Volunteers; about the organisation.	Irish Volunteers; sobre l'organització. 0.802	
WikiMatrix	1.1795002707991602	She said that she liked the Syrian President.	Ella va dir que li agradava el president sirià. 0.600	
WikiMatrix	1.1793223809400522	Once they reach the end they then turn a corner.	(Xoy) Quan arriba a una cresta consecutiva, havent recorregut una vall. 0.496	
WikiMatrix	1.1793216017274057	If this is an emergency, leave a message.	Si això és una emergència, deixem un missatge. 0.846	
WikiMatrix	1.1787946671740444	Your covenant and your signs have been forever.	I tu hi estaràs a mercè de les onades i de les teves associacions.	0.296
WikiMatrix	1.1787169879891684	If they were inspired, why not also the Bible and the Koran?	Si eren inspirats, per què no també la Bíblia i l'Alcorà?	0.840
WikiMatrix	1.1786787802906635	It depends on You - Your efforts. Your determination. Your will to survive!"	Depèn damunt Tu - Els vostres esforcos. La vostra determinació. La vostra voluntat de sobre	

Figura 9: WikiMatrix amb les puntuacions una vegada passat pel Bicleaner

Com l'objectiu era comprovar si la puntuació del Bicleaner era suficientment precisa havia de comparar-la amb una avaluació manual. Per a això vaig decidir fer una mostra de 500 línies de forma aleatòria. Per a això vaig fer servir el següent comandament:

```
cat corpusclassificat.txt | shuf | head -n500 > aleatori500.txt
```

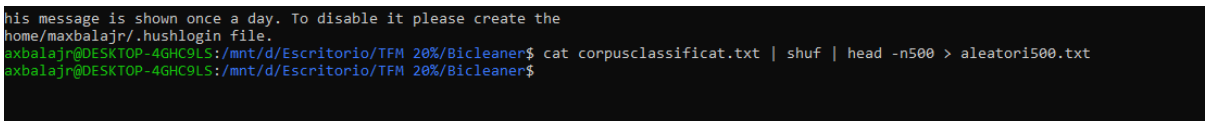


Figura 10. Comandament d'AWK per randomitzar i extreure 500 segments del corpus

Posteriorment, vaig repetir aquesta mateixa operació però amb 200 segments, ja que va ser el nombre de segments que finalment vaig decidir que avaluaria.

Una vegada creada aquesta mostra vaig haver de tornar a canviar la formació de columnes del document amb el comandament d'AWK, ja que l'eina que anava a utilitzar, KEOPS, demanava un document de text amb només dues columnes separades per tabuladors, les dels segments en català i anglès.

```
awk -F "\t"{FS=OFS="\t"} {print $2, $3 > (KEOPS2columnes.txt)}' aleatori500.txt
```

Tot seguit vaig procedir a instal·lar l'eina.

3.4. Avaluació manual amb KEOPS

El KEOPS o *Keen Evaluator Of Parallel Sentences*, com ja he mencionat anteriorment, és una eina que permet fer una avaluació manual de frases paral·leles, tant de forma local com amb *docker*.

Com vaig tenir problemes per instal·lar el programa localment, vaig decidir instal·lar-lo amb *docker*.

En primer lloc, em vaig baixar l'arxiu comprimir de KEOPS[19] i el vaig extreure. Després vaig d'instal·lar el *docker* amb el comandament:

```
sudo apt install docker-compose
```

Posteriorment, vaig executar el comandament que servia per iniciar el *docker* i posar em marxa el servidor:

```
docker-compose up -d
```

Es va produir un error amb Python durant la configuració inicial del KEOPS. Per solucionar-ho, vaig modificar el fitxer "Dockerfile". Tot seguit vaig tornar a intentar la instal·lació, però encara no funcionava. Fallava perquè faltava el fitxer "vendor/autoload.php" i vaig crear-lo.

Tot seguit vaig muntar el KEOPS des de zero amb els següents comandaments:

1. `docker-compose down`
2. `docker system prune -a`
3. `docker-compose up -d`

[19] https://github.com/paracrawl/keops [Accedit el 12 de juliol de 2023]

Per engegar l'allotjador o *host* de KEOPS primerament s'havia d'executar "dockerd" a una finestra que s'havia de deixar oberta perquè l'allotjador funcionés.

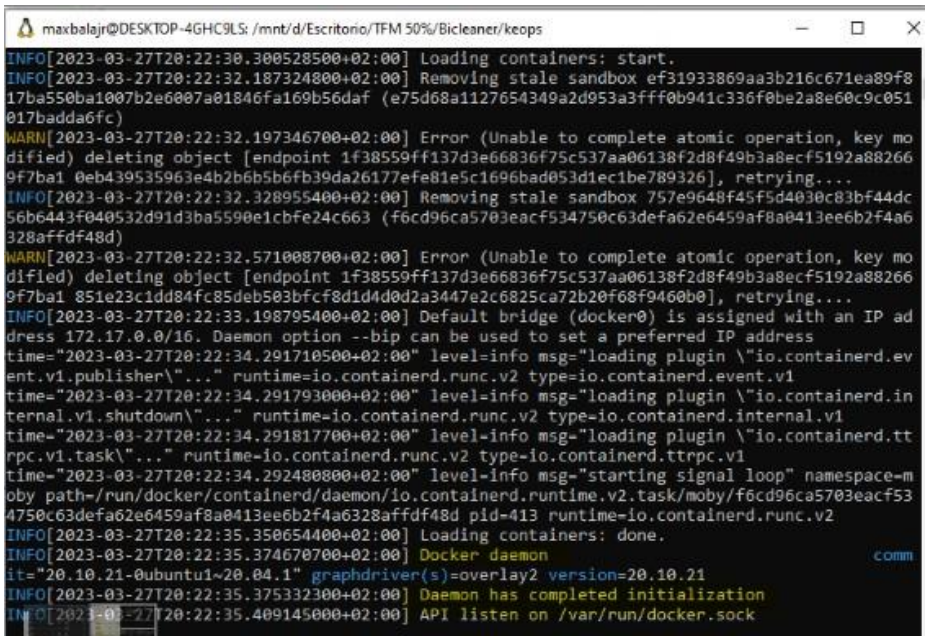


Figura 11. Finestra dockerd

L'allotjador de KEOPS el podem trobar introduint "http://localhost:8080" al nostre navegador, una vegada està tot configurat. Només cal iniciar sessió.

Figura 12. Pantalla de benvinguda a KEOPS

KEOPS està fet per a dos perfils d'usuaris diferents, els avaluadors i els gestors de projectes, amb dues finestres diferents. Per poder avaluar manualment jo sola els segments vaig haver de fer tant de

gestora de projectes com d'avaluadora i enviar-me l'encàrrec a mi mateixa. A la finestra "Tasks" es poden veure les tasques que té assignat l'usuari i en la finestra "Management" les tasques que l'usuari està gestionant com a gestor de projectes. La finestra "Management" inclou, a més, diverses subfinestres per poder gestionar els projectes, els usuaris, els idiomes, les invitacions a projectes, els corpus i els suggeriments.

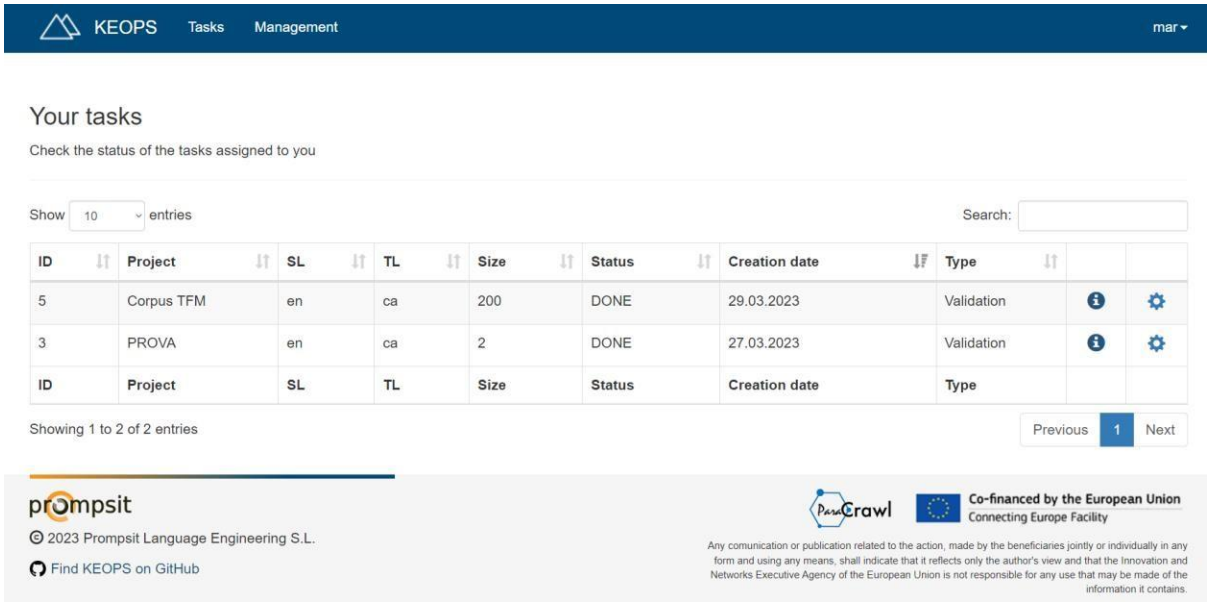


Figura 13. Pantalla de la finestra "Tasks" a KEOPS

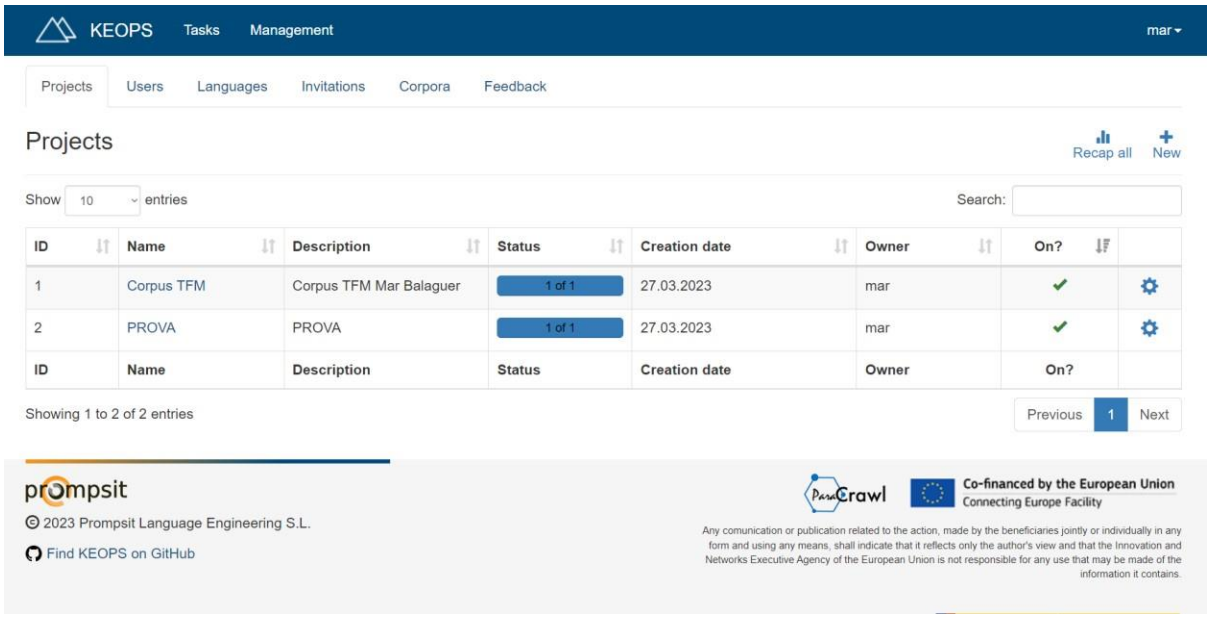


Figura 14. Pantalla de la finestra "Projects" a KEOPS

Per poder començar a avaluar, primer de tot vaig haver d'editar el meu perfil a la finestra Management>Users>mar>Edit user. Vaig haver de posar el meu correu electrònic personal i, el més important, la combinació d'idiomes amb què treballava, que havia de coincidir a la perfecció amb la del corpus. A més, vaig haver de marcar la casella "Active", tal com estava disponible per realitzar la tasca i desar el perfil. Tot seguit vaig haver d'anar a Management>Corpora i pujar el corpus de 200 segments que havia creat anteriorment.

El que vaig haver de fer a continuació va ser crear un nou projecte a dins de KEOPS a Management>Project amb un nom i descripció. A continuació vaig haver de crear una nova tasca a dins del projecte. A l'hora de crear una nova tasca et demanava el mode de tasca, la llengua d'origen, la llengua meta, l'avaluador i el corpus que volia fer servir.

Per a aquest projecte vaig triar el mode *Validation* (validació), ja que el que volia comprovar és si l'alineació era vàlida. Com a *source* vaig posar l'anglès i com a *target*, el català. Com la combinació de llengües coincidia amb la que jo m'havia posat al perfil inicialment, em donava l'opció de triar-me a mi mateixa com a avaluadora. Finalment, vaig seleccionar el corpus que havia pujat anteriorment.

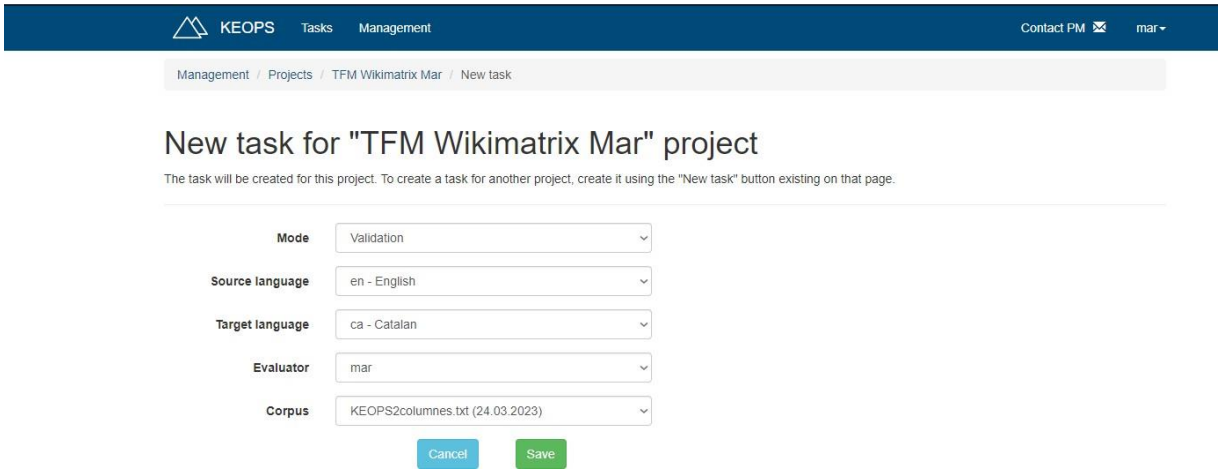


Figura 15. Pantalla de configuració d'una tasca nova a KEOPS

Una vegada fet això, a la pestanya de "Tasks" va aparèixer la tasca que acabava de crear i vaig fer clic al símbol de reproduir de la columna de la dreta per començar l'avaluació manual.

KEOPS el que fa és mostrar el segment en ambdues llengües que has triat i, en el cas del mode *Validation* et dona 7 botons amb diferents opcions, tenint en compte les directrius de validació de l'ELRC (*European Language Resource Coordinatio*n). Les opcions són les següents:

1. *Wrong language id* (identificador de llengua incorrecte)
2. *Incorrect alignment* (alineació incorrecta)
3. *Wrong tokenization* (tokenització incorrecta)
4. *MT translation* (traducció automàtica)
5. *Translation error* (error de traducció)
6. *Free translation* (traducció lliure)
7. *Valid translation* (traducció vàlida)

Figura 16. Pantalla d'avaluació de segments de KEOPS

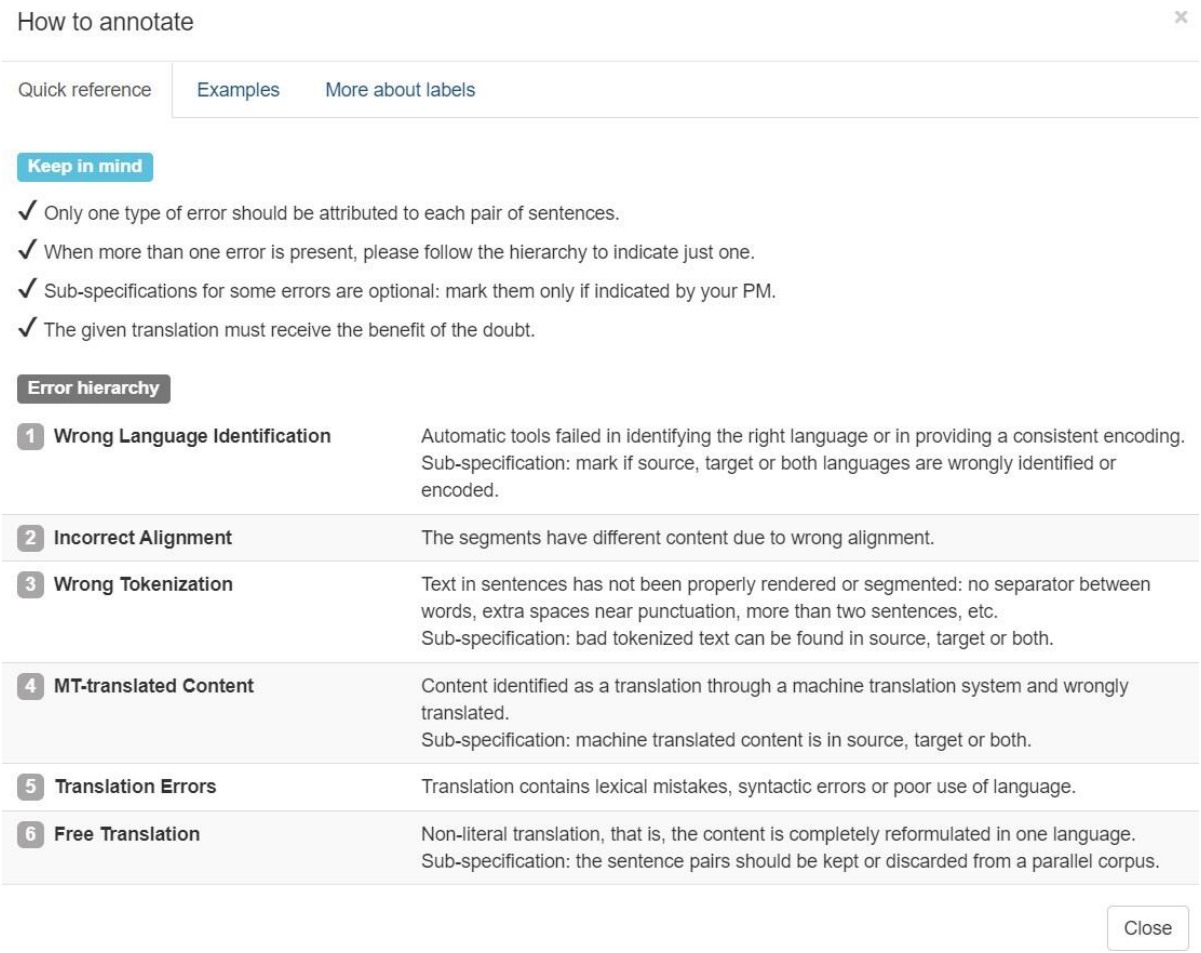
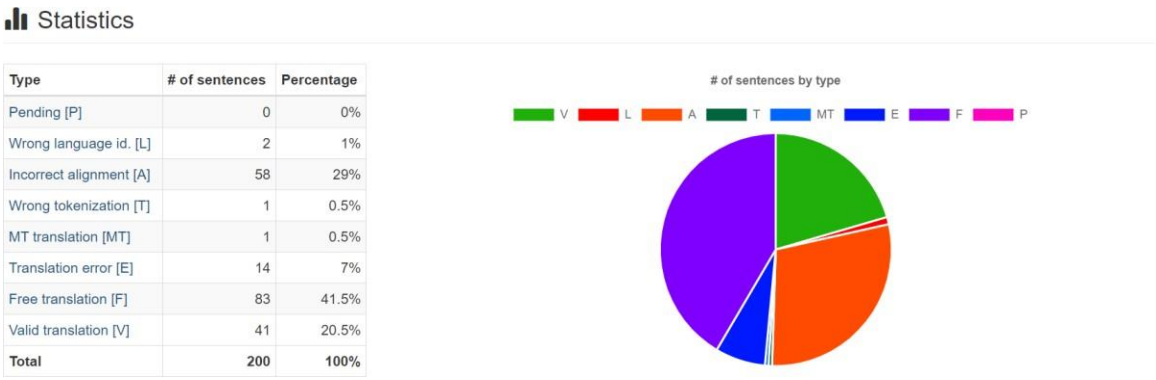


Figura 17. Criteris de validació de l'ELRC al KEOPS

Una vegada vaig avaluar els 200 segments, KEOPS em va permetre veure les estadístiques dels 200 segments en forma de gràfic i em va permetre descarregar dos arxius en format TSV, un del resum i un altre de la puntuació segment per segment.

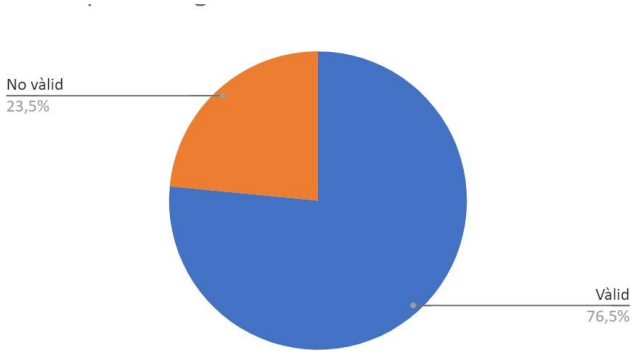


Type	# of sentences	Percentage
Pending [P]	0	0%
Wrong language id. [L]	2	1%
Incorrect alignment [A]	58	29%
Wrong tokenization [T]	1	0.5%
MT translation [MT]	1	0.5%
Translation error [E]	14	7%
Free translation [F]	83	41.5%
Valid translation [V]	41	20.5%
Total	**200**	**100%**

Figura 18. Estadístiques finals després de l'avaluació al KEOPS

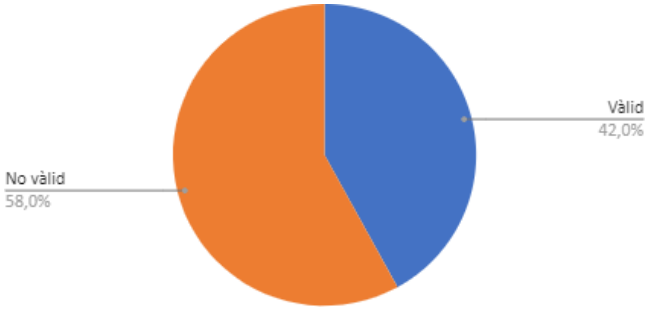
4. Resultats

A l'hora de tenir en compte si els segments puntuats amb el Bicleaner són bons o dolents he tingut en compte dos llindars diferents (Ramírez-Sánchez et al., 2020; Defauw et al., 2019). Com el Bicleaner puntua entre 0 i 1, el primer llindar seria el de 0,5. Aquest llindar representa una qualitat decent, però no excel·lent. El segon llindar l'he establert a 0,7, que representa ja una qualitat bona, amb la filtració de la majoria dels segments amb errors. No obstant això, per facilitar la gestió de les dades al document Excel, vaig passar les xifres a milers, així que d'ara en avant parlaré del llindar de 500 (0,5) i del llindar de 700 (0,7). Els segments amb puntuacions per sobre dels llindars es consideraran "Vàlids" i els que estan per sota "No vàlids", amb l'objectiu de poder comparar aquestes dades amb el KEOPS posteriorment.



Com podem observar al gràfic, si establim el llindar en 500, un 76,5% (153 segments) dels segments són, segons el Bicleaner, segments aprofitables. Només un 23,5% (47 segments) estaria per sota dels 500 punts.

Figura 19. Recompte de segments vàlids del Bicleaner amb el llindar a 500



Si establim el llindar en 700 els percentatges canvien una mica, ja que ara el percentatge de segments vàlids passa de ser abans un 76,5% a ser un 42% (84 segments). Els segments no vàlids pugen fins a un 58%.

Figura 20. Recompte de segments vàlids del Bicleaner amb el llindar a 700

Per observar amb més detall com es reparteixen les puntuacions del Bicleaner en ambdós llindars tenim el següent histograma:

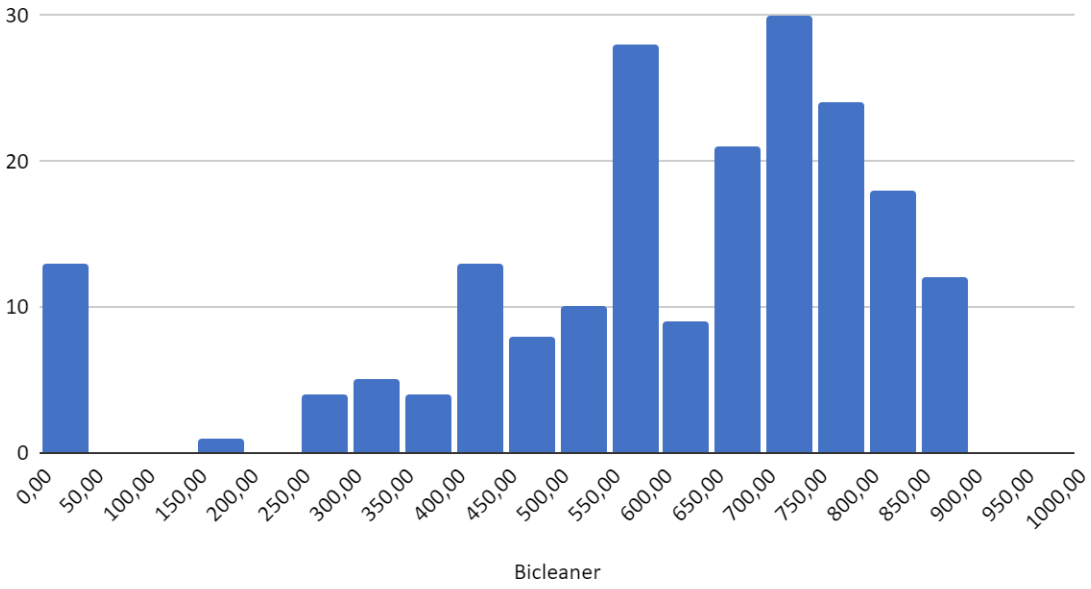
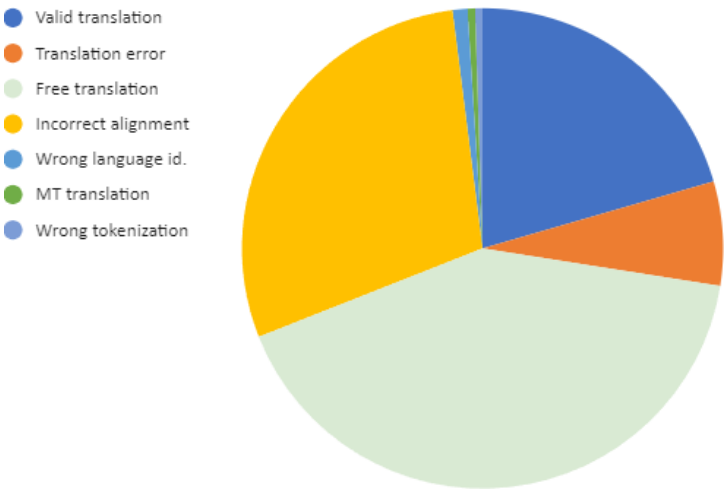


Figura 21. Histograma del Bicleaner

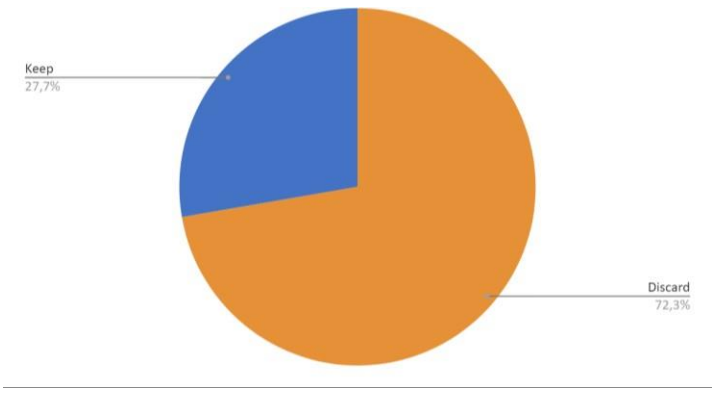
Com podem observar, el gruix de les puntuacions està entre els 650 i els 850 punts, amb un pic als 550. Això vol dir que, segons el Bicleaner, la gran majoria de les parelles de segments s'assemblen entre elles i, mirant aquestes dades, podríem considerar que WikiMatrix és un corpus de bona qualitat. Tot i això, s'observa clarament que no hi ha cap segment de 950-1000, és a dir, que no considera que cap parell sigui una traducció exactament igual. També trobem un buit entre 50 i 250.



Per comprovar si aquestes dades són correctes les contrastarem amb l'avaluació manual realitzada amb el KEOPS. Com s'ha explicat anteriorment, el KEOPS no avalua amb una puntuació numèrica, sinó que li assigna una categoria a cada segment, depenent del tipus d'error o de la falta d'errors.

Figura 22. Percentatges de cada error a KEOPS

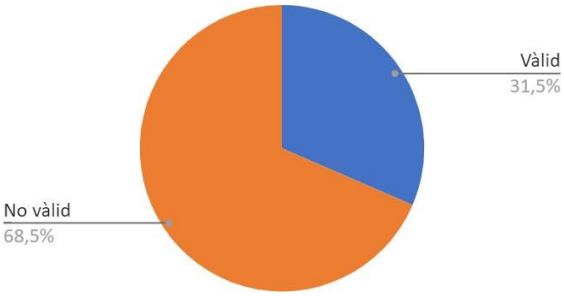
Si analitzem el gràfic, podem veure que la gran majoria, un 41,5% dels segments, són traduccions lliures (*Free translation*). D'aquestes traduccions lliures n'hem mantingut un 27,7% i n'hem descartat un 72,3%. La segona categoria amb més nombre de segments és la d'alineació incorrecta (Incorrect alignment) amb un 29%. En tercer lloc, tenim les traduccions vàlides (*Valid translation*) amb només un 20,5% del total, seguides dels errors de traducció (*Translation error*), un 7% i amb només un 1% tenim l'error d'identificador d'idioma incorrecte (*Wrong language id*). Per acabar, amb només un 0,5% tenim TA (MT) i tokenitzacions incorrectes (*Wrong tokenization*).



De 83 segments de traducció lliure s'ha considerat que només 23 (27,7%) eren traduccions correctes. Si sumem això a les traduccions vàlides, 41 segments (72,3%), trobem que de 200 segments de la mostra, segons l'avaluació manual, només 64 segments són vàlids.

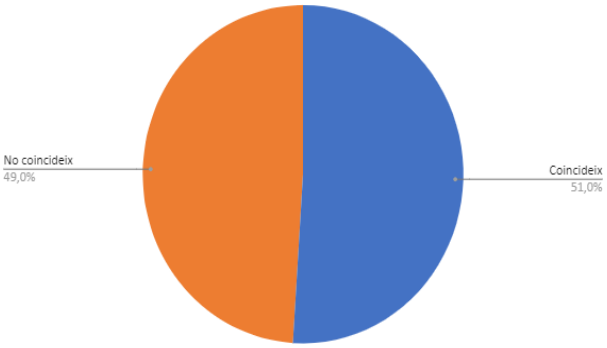
Figura 23. Percentatge de segments que s'han mantingut o descartat de la categoria Free Translation

Per poder comparar les dades del Bicleaner i del KEOPS hem d'unificar el tipus de dades, ja que no es poden comparar puntuacions numèriques amb noms de categories. Per aconseguir-ho he assignat els segments amb bona puntuació del Bicleaner i els de traduccions vàlides i lliures que he decidit mantenir (Fr*ee translation: keep*) del KEOPS com a "Vàlid". Per contra, els segments amb les puntuacions del Bicleaner per sota dels llindars establerts prèviament i els pertanyents a la resta de categories del KEOPS són "No vàlid".



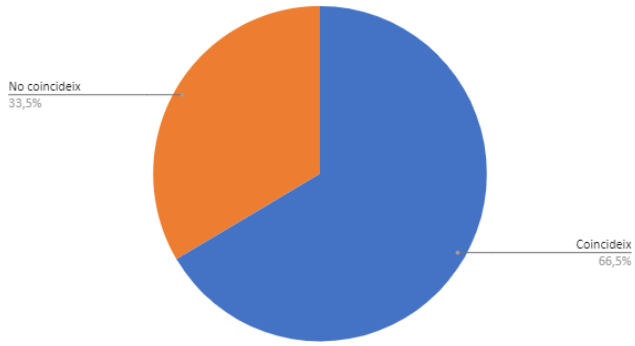
Tenint en compte aquesta classificació, el gràfic ens mostra que només un 31,5% dels 200 segments de la mostra són vàlids, amb un 68,5% de segments no vàlids.

Figura 24. Percentatge de segments vàlids al KEOPS



Si comparem les dades del KEOPS amb les del Bicleaner amb el llindar al 500 només un 49% dels segments coincideixen. És a dir, en un 51% dels casos, els segments tenen la mateixa classificació de Vàlid/No Vàlid tant a Bicleaner com a KEOPS.

Figura 25. Percentatge de coincidències entre els segments vàlids del Bicleaner amb el llindar 500 i els del KEOPS



Si augmentem el llindar a 700, també augmenten les coincidències (66,5%), perquè es filtren les opcions amb més errors. Això ens deixa amb 33,5% de segments en què la classificació no coincideix en ambdues eines.

Figura 26. Percentatge de coincidències entre els segments vàlids del Bicleaner amb el llindar 500 i els del KEOPS

Per poder comprovar, a més, quins tipus d'error del KEOPS penalitzaven més al Bicleaner, s'ha extret la mitjana de puntuacions de totes les categories.

Valid translation (41)	698
Free translation: keep (23)	695
Wrong tokenization (2)	660
Free translation: discard (59)	633
Translation error (14)	592
Incorrect alignment (58)	491
Machine translation (1)	0
Wrong language Id. (2)	0

Figura 27. Mitjana de les puntuacions del Bicleaner per categoria

Els errors que més penalitzen són *Machine translation* i *Wrong language id* amb un 0, seguits d'*Incorrect alignment* amb una mitjana inferior al llindar dels 500. Els *Translation errors* no reben gaire penalització, ja que la mitjana dels segments marcats amb aquest error és de 592, per sobre dels 500. La mitjana de *Free translation: discard* resulta molt similar a la de *Free translation: keep*, cosa que no hauria de poder ser, ja que la segona categoria són segments que s'han descartat per la quantitat

d'errors que contenen. La categoria d'error que menys penalitza, amb diferència, és la de *Wrong tokenization*, amb una mitjana de 660.

5. Conclusió

Després de veure els resultats que s'acaben de presentar, se'n poden extreure algunes conclusions. No obstant això, cal remarcar que la prova de l'eina que s'ha fet ha sigut amb un mostreig de només 200 segments d'un sol corpus, per la qual cosa no són resultats concloents, sinó que s'haurien de contrastar amb altres corpus de millor qualitat o altres combinacions lingüístiques.

Dit això, com a punts positius podem dir que Bicleaner és una eina força fàcil d'utilitzar. Sí que és cert que cal un nivell mínim de coneixements per executar els *scripts* i modificar el document del corpus, però no cal un nivell tan alt com amb altres eines. Es basa en la distància d'edició, que és una de les mètriques més conegudes i emprades a l'hora d'avaluar corpus i traduccions de motors de traducció automàtica. A més, el percentatge mínim de coincidència amb l'avaluació humana és del 49%, cosa que indica que, almenys en la meitat dels casos, encerta. No només això, sinó que també s'ha de tenir en compte que Bicleaner compta amb una versió millorada, el Bicleaner AI, que segons els estudis (Zaragoza-Bernabeu et al., 2022) dona millors resultats i s'hauria d'aprofundir en aquesta opció. Per desgràcia no s'ha pogut incloure en aquest treball per problemes amb les dependències a Python i manca de temps. A més, tampoc s'ha pogut entrenar un motor de TA amb el corpus "brut" i un altre amb el corpus filtrat i utilitzar els resultats per veure si hi ha cap variació en els resultats.

No obstant això, si el que cerquem és una avaluació més acurada, la coincidència és massa baixa per poder confiar-hi plenament, ja que el fet que només coincideixi en un 50-65% dels casos representa que, com a mínim, un 40% dels segments tenen una puntuació que no representa la qualitat "real". A més, tot i que encerta molt bé les puntuacions de la part mitjana-alta del rang total (400-900), li costa avaluar de forma acurada els segments amb errors i, si veu algun error, o hi posa un 0 directament o puja als 250. Pel que fa a les traduccions vàlides, també li costa detectar-les: en aquest mostreig a partir dels 900 punts no n'ha detectat cap. Cal mencionar, però, que s'hauria d'ampliar el nombre de mostres per poder fer-ne una avaluació més exhaustiva.

Emprada amb una segona avaluació com a suport, Bicleaner ajuda a fer-se una idea de la qualitat general del corpus o de la traducció, però per si sola no és fiable.

Per concloure, és important afegir que l'objectiu final d'aquest TFM és obrir camí a futurs estudis d'avaluació de qualitat i tractament de corpus bilingües, ja que és un tema extremadament útil amb una forta mancança de recerca. Espero que serveixi a futures estudiants per inspirar-les a continuar per on hem començat.

6. Bibliografia

Abercrombie, D. (1965). *Studies in phonetics and linguistics* (No. 10). Oxford University Press.

Artetxe, M., Aldabe, I., Agerri, R., Perez-de-Viñaspre, O., i Soroa, A. (2022). *Does Corpus Quality Really Matter for Low-Resource Languages?* (arXiv:2203.08111). arXiv. https://doi.org/10.48550/arXiv.2203.08111

Banerjee, S., i Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65-72. https://aclanthology.org/W05-0909 [Consultat el 22 de juny de 2023]

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., i Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4555-4567. https://doi.org/10.18653/v1/2020.acl-main.417

Briakou, E., i Carpuat, M. (2020). Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1563-1580. https://doi.org/10.18653/v1/2020.emnlp-main.121

Briakou, E., i Carpuat, M. (2021). Beyond Noise: Mitigating the Impact of Fine-grained Semantic Divergences on Neural Machine Translation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7236-7249. https://doi.org/10.18653/v1/2021.acl-long.562

Briakou, E., i Carpuat, M. (2022a). *Can Synthetic Translations Improve Bitext Quality?* (arXiv:2203.07643). arXiv. https://doi.org/10.48550/arXiv.2203.07643

Briakou, E., Wang, S. I., Zettlemoyer, L., i Ghazvininejad, M. (2022b). *BitextEdit: Automatic Bitext Editing for Improved Low-Resource Machine Translation* (arXiv:2111.06787). arXiv. https://doi.org/10.48550/arXiv.2111.06787

Briva-Iglesias, V. (2020). *Traducción automática inglés-catalán: tecnología de vanguardia, calidad y productividad.* [TFM]. Universitat Autònoma de Barcelona

Busa, R. (1951). Rapida E Meccanica Composizione E Pubblicazione Di Indici E Concordanze Di Parole Mediante Macchine Elettrocontabili. *Aevum*, *25*(6), 479-493.

Chaudhary, V., Tang, Y., Guzmán, F., Schwenk, H., i Koehn, P. (2019). Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 261-266. https://doi.org/10.18653/v1/W19-5435

Dash, N. S., i Arulmozi, S. (2018). *History, features, and typology of language corpora*. Springer.

Defauw, A., Vanallemeersch, T., Szoc, S., Everaert, F., Van Winckel, K., Scholte, K., Brabers, J., i Van den Bogaert, J. (2019). Collecting domain specific data for MT: An evaluation of the ParaCrawlpipeline. *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, 186-195. https://aclanthology.org/W19-6733 [Consultat el 22 de juny de 2023]

Doddington, G. (2002). Automatic evaluation of machine translation quality using ngram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, 138. https://doi.org/10.3115/1289189.1289273

EAGLES (Expert Advisory Group on language Engineering Standards) (1996). Text Corpora Working Group Reading Guide. *EAGLES Document EAG-TCWG-FR-2*

Esplà, M., Forcada, M., Ramírez-Sánchez, G., i Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, 118-119. https://aclanthology.org/W19-6721 [Consultat el 22 de juny de 2023]

Esteban, S. (2020) *Evaluación de la calidad de la traducción de motores de traducción automática neuronal en textos del ámbito jurídico.* [TFM]. Universitat Autònoma de Barcelona

Forcada, M. L., Sánchez Martínez, F., i Pérez Ortiz, J. A. (2021). *Manual d'informàtica i de tecnologies per a la traducció*. Universitat d'Alacant. http://rua.ua.es/dspace/handle/10045/53085 [Consultat el 22 de juny de 2023]

Francis, N. W. (1982). Problems of assembling and computerizing large corpora. In S. Johansson (ed.), Computer Corpora in English Language Research (pp. 7-24). Norwegian Computing Centre for the Humanities.

Francis, W. (1992). Language corpora B.C.. In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991* (pp. 17-32). De Gruyter Mouton. https://doi.org/10.1515/9783110867275.17

He, Y., i Way, A. (2009). Learning Labelled Dependencies in Machine Translation Evaluation. *Proceedings of the 13th Annual Conference of the EAMT, 44-51. https://aclanthology.org/2009.eamt-1.7.pdf* [Consultat el 22 de juny de 2023]

Juilland, A. i Chang Rodríguez, E. (1964). *Frequency Dictionary of Spanish Words*, Mouton

Khayrallah, H., i Koehn, P. (2018). On the Impact of Various Types of Noise on Neural Machine Translation. *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 74-83. https://doi.org/10.18653/v1/W18-2709

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P. J., i Guzmán, F. (2020). Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation* (pp. 726-742). https://aclanthology.org/2020.wmt-1.78 [Consultat el 22 de juny de 2023]

Koehn, P., Guzmán, F., Chaudhary, V., i Pino, J. (2019). Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 54-72. https://doi.org/10.18653/v1/W19-5404

Koehn, P., i Knowles, R. (2017). Six Challenges for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, 28-39. https://doi.org/10.18653/v1/W17-3204

Koehn, P., Khayrallah, H., Heafield, K., i Forcada, M. L. (2018). Findings of the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the third conference on machine translation: shared task papers* (pp. 726-739). https://doi.org/10.18653/v1/W18-64080

Lample, G., Ott, M., Conneau, A., Denoyer, L., i Ranzato, M. (2018). Phrase-Based & Neural Unsupervised Machine Translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5039-5049. https://doi.org/10.18653/v1/D18-1549

Laskar, S. R., Khilji, A. F. U. R., Pakray, P., i Bandyopadhyay, S. (2020). EnAsCorp1.0: English-Assamese Corpus. *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, 62-68. https://aclanthology.org/2020.loresmt-1.9 [Consultat el 22 de juny de 2023]

Lavie, A. (2011). Evaluating the Output of Machine Translation Systems. *Proceedings of Machine Translation Summit XIII: Tutorial Abstracts*. https://aclanthology.org/2011.mtsummit-tutorials.3 [Consultat el 22 de juny de 2023]

Llamazares, M. V. (2008). Lingüística con corpus (I). *Estudios humanísticos. Filología*, (30), 329-349. https://dialnet.unirioja.es/descarga/articulo/3332675.pdf [Consultat el 22 de juny de 2023]

McEnery, T., i Hardie, A. (2013). The history of corpus linguistics. Oxford University Press

McEnery, T., Xiao, R., i Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Moorkens, J., Castilho, S., Gaspari, F., i Doherty, S. (2018). Translation Quality Assessment: From Principles to Practice (Vol. 1). Springer International Publishing. https://doi.org/10.1007/978-3-319-91241-7

Moreno, S. (2022). *Creació i avaluació d'un motor de TAN català-japonès [TFM] Universitat Autònoma de Barcelona*

Oliver González, A. (2014). Traducción y tecnologías: procesos, herramientas y recursos, septiembre 2014. Universitat Oberta de Catalunya. http://hdl.handle.net/10609/79008 [Consultat el 22 de juny de 2023]

Orăsan, C., Ha, L. A., Evans, R., Hasler, L., i Mitkov, R. (2007). Corpora for Computational Linguistics. *Ilha do Desterro: A Journal of English Language, Literatures in English and Cultural Studies*, (52), 65-101. https://www.redalyc.org/articulo.oa?id=478348690003 [Consultat el 22 de juny de 2023]

Papavassiliou, V., Sofianopoulos, S., Prokopidis, P., i Piperidis, S. (2018). The ilsp/arc submission to the wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 928-933). https://doi.org/10.18653/v1/W18-6484

Papineni, K., Roukos, S., Ward, T., i Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. https://doi.org/10.3115/1073083.1073135

Pestov, I. (2018). *A history of machine translation from the Cold War to deep learning*. FreeCodeCamp.Org. https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/[Consultat el 22 de juny de 2023]

Ramírez-Sánchez, G., Zaragoza-Bernabeu, J., Bañón, M., i Rojas, S. O. (2020). Bifixer and Bicleaner: Two open-source tools to clean your parallel data. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 291-298. https://aclanthology.org/2020.eamt-1.31 [Consultat el 22 de juny de 2023]

Rojo, G. (2021). *Introducción a la lingüística de corpus en español*. Routledge.

Rossenbach, N., Rosendahl, J., Kim, Y., Graça, M., Gokrani, A., i Ney, H. (2018). The RWTH Aachen University Filtering System for the WMT 2018 Parallel Corpus Filtering Task. Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 946-954. https://doi.org/10.18653/v1/W18-6487

Sánchez, Mª del M. i Rico, C. (2020)*Traducción automática. Conceptos clave, procesos de evaluación y técnicas de posedición*. Comares.

Sánchez-Cartagena, V. M., Bañón, M., Ortiz-Rojas, S., i Ramírez, G. (2018). Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 955-962. https://doi.org/10.18653/v1/W18-6488

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., i Guzmán, F. (2021a). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 1351-1361. https://doi.org/10.18653/v1/2021.eacl-main.115

Schwenk, H., Wenzek, G., Edunov, S., Grave, E., Joulin, A., i Fan, A. (2021b). CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 6490-6500. https://doi.org/10.18653/v1/2021.acl-long.507

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., i Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, 223-231. https://aclanthology.org/2006.amta-papers.25 [Consultat el 22 de juny de 2023]

Stanojevic, M., i Sima'an, K. (2015). Evaluating MT systems with BEER. *The Prague Bulletin of Mathematical Linguistics*, 104, 17-26. https://doi.org/10.1515/pralin-2015-0010

Torruella, J., i Llisterri, J. (1999). Diseño de corpus textuales y orales. *Filología e informática. Nuevas tecnologías en los estudios filológicos*, 45-77. Seminari de Filologia i Informàtica, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona. https://joaquimllisterri.cat/publicacions/Torruella_Llisterri_99.pdf [Consultat el 22 de juny de 2023]

Ye-Yi Wang, Acero, A., i Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding* ,577-582. https://doi.org/10.1109/ASRU.2003.1318504

Zaragoza-Bernabeu, J., Ramírez-Sánchez, G., Bañón, M., i Ortiz Rojas, S. (2022). Bicleaner AI: Bicleaner Goes Neural. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 824-831. https://aclanthology.org/2022.lrec-1.87 [Consultat el 22 de juny de 2023]

7. Annex: Taula amb els segments inclosos en l'avaluació i les seves puntuacions

Anglès	Català	Bicleaner	KEOPS
These cases are representative of a system of aristocratic patronage in which scientists received funding to develop either immediately useful things or to entertain.	Són casos representatius d'un sistema de mecenatge aristocràtic en el què els científics rebien finançament per desenvolupar de manera immediata invents útils i per a entretenir.	696	Valid translation
He died in April 2009 in Tromsø, Norway.	Va morir el 12 de desembre de 2008 a Tromsø, Noruega.	654	Translation error
The Padrón Real (Spanish pronunciation: , Royal Register), known after 2 August 1527 as the Padrón General (Spanish: , General Register), was the official and secret Spanish master map used as a template for the maps present on all Spanish ships during the 16th century.	El Padrón Real (conegut a partir del 2 d'agost de 1527 com a Padrón General,) era el mapa espanyol principal, oficial i secret, que era utilitzat com a model per als mapes i les cartes nàutiques presents a tots els vaixells espanyols durant els anys 1500.	762	Free translation [discard]
Besides this, there are only small-scale factories and businesses.	A més hi ha unes 160 empreses de petita indústria i serveis.	540	Free translation[discard]
The water of the river is crystal clear and pure.	Les aigües en la desembocadura són netes i cristal·lines.	652	Free translation[keep]
This time around, they're centre stage.	A vegades, aquestes estan entremesclades en la part central.	560	Free translation[discard]
William Tell stands apart from the general merriment, however: he is consumed with ennui at Switzerland's continued oppression (Il chante, et l'Helvétie pleure sa liberté – "He sings, and Helvetia mourns her liberty").	Guillem Tell roman apartat de l'alegria general, ja està consumit amb ennui, per la contínua opressió de Suïssa (Il chante, et l'Helvétie pleure sa liberté - Ell canta, i Suïssa plora la seva llibertat).	832	Valid translation
That is, equilibrium is attained.	En aquest punt hi ha equilibri.	562	Free translation[discard]
Other individual factors such as gender, age, trust and personality, may also account for the paradox.	Un gran nombre de factors, com el gènere, l'edat, l'estatus social, la riquesa i la professió, determinaven el pentinat.	536	Translation error
Other principal cast members include Dave Franco, Analeigh Tipton, and Cory Hardrict.	També apareixen Dave Franco, Analeigh Tipton, i Cory Hardrict.	858	Free translation[discard]
Tham Luang cave rescue "Onemi confirma a 33 mineros atrapados en yacimiento en Atacama".	Rescat de la cova de Tham Luang Rescat al pou de Totalán «Un vídeo mostra els miners atrapats a Xile prims però animats».	0	Free translation[discard]
Zipolite (in Spanish).	«Zippeite» (en anglès).	738	Translation error
Future bloc calls on premier to immediately resign.	«Future bloc calls on premier to immediately resign» (en anglès).	750	Free translation[discard]
After not accomplishing it, the draft was leaked, based on information extracted from the internet.	Després de no aconseguir-ho es va filtrar l'esborrany, basat en informació treta d'internet.	820	Valid translation
He also compared the structure of the heart in other vertebrates and annelids.	Igualment comparà l'estructura del cor en altres vertebrats i anèl·lids.	816	Valid translation
The Coulter pine occurs in a number of forest plant associations; for example, At higher elevations forestation of the San Jacinto Mountains Coulter Pine is co-dominant with the California black oak.	Apareix en una sèrie de formacions forestals; per exemple, a majors altures de les muntanyes San Jacinto aquest pi és dominant juntament amb el Quercus kelloggii.	858	Free translation[discard]
The proof is written as a series of lines in two columns.	La demostració s'escriu com una sèrie de línies en dues columnes.	774	Valid translation
The album was released on 22 September 2004, the same day that John Norum became a father.	L'àlbum es va començar a vendre el 22 de setembre de 2004, el mateix dia que John Norum va ser pare.	658	Free translation[keep]
In the past the area of the district was covered with deep forest.	Una àrea del Janícul estava coberta de boscos sagrats.	424	Free translation[discard]

The trail is very rocky and goes by several large boulders.	Aquesta llum és molt difusa i pren uns colors groguencs.	770	Incorrect alignment
He was praetor in 148 BC, and he received the command in Macedonia.	L'any 148 aC va ser pretor i va rebre Macedònia com a província.	756	Free translation[discard]
Geology : an introduction to physical geology.	Geology: An Introduction to Physical Geology (en anglès).	672	Free translation[discard]
Other problems include hybridization or genetic pollution with introduced, non-native trouts, illegal fishing and poor water and fisheries management.	Altres factors són erupcions volcàniques, inundació del niu o pesca local il·legal.	388	Free translation[discard]
In 2008, users of the imageboard 4chan posted numerous sexually explicit parodies and cartoons illustrating Rule 34.	El 2008, els usuaris de la tauler d'imatges 4chan publicaren nombroses paròdies sexualment explícites i dibuixos animats il·lustrant la Regla 34.	704	Free translation[keep]
General meetings of work collectives at every enterprise and office elected the committees for tenures of two and one-half years.	Els cònsols majors i menors de cada parròquia eren elegits pels consells de comú per a un període de 2 anys.	404	Incorrect alignment
In 1944, the Saskatchewan CCF formed the first socialist government in North America.	El 1944, el CCF de Saskatchewan va formar el primer govern socialista a Amèrica del Nord.	814	Valid translation
From there, the group aims for Italy, dreaming of making the aliyah to Palestine to take part in the Zionist project of reclaiming a Jewish homeland.	A partir de la segona diàspora, les comunitats religioses mantindran el somni d'una tornada a Palestina per recrear un Estat jueu.	678	Free translation[discard]
Sweet gelatin pudding spiced with saffron and cardamom.	Limita amb Gilet i Sagunt.	0	Incorrect alignment
Copper from Hungary was transported through Antwerp to Lisbon, and from there shipped to India.	El coure d'Eslovàquia era transportat a través d'Anvers a Lisboa, i d'allí era enviat a l'Índia.	368	Translation error[discard]
The French troops in the army were then placed under the command of generals de Broglie and Coigny, who were made Marshal.	Les tropes franceses de l'exèrcit foren llavors posades sota el comandament dels generals de Broglie i Coigny, que foren ascendits a mariscals.	832	Valid translation
It was estimated by the BBC that about 3 million people turned up around the Thames to watch the fireworks display.	Avui dia, s'estima que hi ha 13 milions de mormons arreu del món, gràcies a l'activitat missionera.	262	Incorrect alignment
Search.	«Recerques.	0	Valid translation
In September, 2011, Reid's wife was diagnosed with stage 2 breast cancer.	El dia de la mare de 2013 la mare de Alexis va ser diagnosticada amb l'etapa 4 del càncer rectal.	418	Incorrect alignment
Bresimo borders the following municipalities: Ulten, Rumo, Rabbi, Livo, Cis, Malè, and Caldes.	Limita amb els municipis de Caldes, Cis, Livo, Malè, Rabbi, Rumo i Ulten (BZ)	752	Free translation[keep]
The post office closed in 1964, and reopened in 1965.	L'oficina va tancar el 1964 i reobrí el 1965	778	Valid translation
He studied at Transylvania University in Lexington, Kentucky, from which he graduated in 1810.	Després va continuar els estudis a la universitat de Transylvania a Lexington, Kentucky. on es va graduar en dret.	728	Free translation[discard]
Near the end of his life, Bohm began to experience a recurrence of the depression that he had suffered earlier in life.	Això feia sospitar que Everett s'estava començant a recuperar de les males experiències que havia viscut.	696	Incorrect alignment
Two hundred Jews were awaiting the same fate, in spite of their innocence, and the whole Jewish community had been fined 4,000,000 piastres.	Prop de 100.000 persones jueves van ser exiliades, de manera que la seva població jueva gairebé va ser eliminada per complet.	412	Incorrect alignment
Before its release to news media, congressional staff digitally added into this 2013 official portrait the heads of four members absent in the original photo shoot.	Abans de la seva difusió als mitjans de comunicació, personal del Congrés photoshopped en aquest 2013 el retrat oficial dels caps dels quatre membres absents de la sessió de fotos originals.	682	Translation error
"Obama, Cameron 'no doubt' Syrian regime in chem attack"".	French, Cameron «No softies in Canada's Pillow Fight League» (en anglès).	848	Incorrect alignment
What has ensued so far has been a remarkable start in the comedy field thanks to a perspective that the world has never seen before.	El seu afany innovador li ha facilitat la troballa d'inspiració en llocs que ningú ha vist abans..	564	Free translation[discard]
Before returning to Damascus, the raiders seized the cave castle of Habis Jaldak in the Yarmuk Valley from its weak Frankish garrison.	Abans de tornar a Damasc, capturaren el castell d'Habis Jaldak, situat en la vall del Yarmuk, que estava defensat per una feble guarnició de croats	618	Free translation[keep]

In 1973, none of the 4 enrolled members lived on the reservation.	En 1973 cap dels 4 membres registrats de la tribu vivia a la reserva.	632	Valid translation
When the solution is shaken, it turns from light blue to a redish color.	Quan s'agrega l'aigua, la barreja canvia a un color blanquinós.	738	Incorrect alignment
McDonagh wrote and directed his drama-dark comedy film Three Billboards Outside Ebbing, Missouri (2017), starring Frances McDormand, Woody Harrelson, and Sam Rockwell, which premiered at the Venice Film Festival on 4 September and won the People's Choice Award at the Toronto International Film Festival on 17 September 2017.	McDonagh va escriure i dirigir la pel·lícula de comèdia negra Three Billboards Outside Ebbing, Missouri (2017), protagonitzada per Frances McDormand, Woody Harrelson i Sam Rockwell, que es va estrenar al Festival de Venècia el 4 de setembre i va guanyar el Premi People's Choice en el Festival Internacional de Cinema de Toronto .	824	Valid translation
The patriarch and his associates manifested consummate prudence in dealing with this matter.	Les associacions conservadores i defensores del patriarcat han lluitat contra la influència d'aquest llibre.	720	Incorrect alignment
The first currency denominated in dollars to be issued in Barbados was in the form of private banknotes introduced in 1882.	La primera vegada que va aparèixer el dòlar a l'Hondures Britànica fou en forma de bitllets el 1855.	594	Incorrect alignment
A second mosaic from Noviomagus may be seen at Fishbourne Roman Palace.	El lloc on es troba el conjunt d'excavacions es diu Palau romà de Fishbourne.	590	Free translation[discard]
With this is joined, where needed and possible, the apostolate of union with Rome.	Va coincidir, més o menys, amb la segona part i esclat del romanticisme a Europa.	718	Incorrect alignment
Greenblatt, along with Tom Knight and Stewart Nelson, co-wrote the Incompatible Timesharing System (ITS), a highly influential time-sharing operating system for the PDP-6 and PDP-10 used at MIT.	Al MIT també va treballar amb Tom Knight i Stewart Nelson per a desenvolupar l'Incompatible Timesharing System en què desenvolupaven un sistema de temps compartit que permetia als usuaris utilitzar diversos programes a la vegada.	732	Free translation[discard]
These oscillations show a phase shift of π, known as Berry's phase.	La cerfluorita és una fluorita que conté ceri, coneguda com a fase sintètica.	582	Incorrect alignment
The colonists arrived ill-prepared to become self-sufficient.	Els àrabs no s'hi van poder establir sòlidament sota els omeies.	294	Incorrect alignment
Sociologist G. S. Ghurye declined the honour as he felt he deserved the higher Padma Vibhushan, given the calibre of others who had received the more prestigious decoration.	Curiositat: Godric Gryffindor només acollia els mags valents i honorables, ja que era el que ell més apreciava.	320	Incorrect alignment
Milton's Paradise Lost, part I, 1981.	2 Museu del Montgrí i del Baix Ter, 1983.	602	Incorrect alignment
The Capital Grille.	La capital és Jerada.	532	Translation error
However, many Caltech faculty members within physics and aeronautical engineering did not view the crisis as a way to strengthen their programs.	No obstant això, altres facultats com la de física, informàtica o meteorologia no realitzen preselecció d'estudiants.	490	Incorrect alignment
The scientific team, which included the future Australasian Antarctic Expedition leader Douglas Mawson, carried out extensive geological, zoological and meteorological work.	L'equip científic, que comptava amb el líder de la futura Expedició Terra Nova, Douglas Mawson, va fer amplis treballs geològics, zoològics i meteorològics.	732	Free translation[keep]
He eulogizes her as a faultless wife and mother of his three children, and describes how she one day requested a rare apple when she was ill.	La descriu com una esposa perfecta i mare dels seus tres fills, i descriu com una vegada va demanar una poma rara quan estava malalta.	582	Valid translation
Eugenia (in Spanish).	«Eugenite» (en anglès).	736	Free translation[discard]
It was offered in three trims, standard, Coronado (luxury) and GT (sport).	Inicialment van oferir-se 3 paquets d'equipament: GL (base), LX (luxòs) i SE (esportiu).	474	Free translation[discard]
وی او اے.	V - Pà.	0	Wrong language id.
Separate forums exist for each model of phones manufactured by Google, Sony, HTC, Samsung, LG Electronics, Motorola, and many others.	Independent del fòrum existeix per a cada model dels telèfons fabricats per Sony, HTC, Samsung, LG, Motorola, i molts altres.	828	Translation error
By the time of the Descent from the Cross and Durán Madonna, van der Weyden has already worked out a far more complex and effective means of mixing temporal and non-temporal effects".	En l'època del Davallament i de la Madonna Durán, Van der Weyden ja havia realitzat simbolismes i significats més complexos en barrejar efectes temporals i intemporals".	874	Free translation[keep]

In some cases he flew through the debris of exploding enemy aircraft, and on one occasion collided with his target, which he claimed as a "probable" victory.	En alguns casos volà entre les restes d'un avió enemic explotant, i en una ocasió topà amb el seu objectiu, reclamant una "victòria probable".	866	Valid translation
I could make it for myself, but also it's a Rapture song.	Segur és que nasqué d'una llegenda i també d'una cançó.	710	Incorrect alignment
Sertularella polyzonias	Vegeu Polarimetria	0	Incorrect alignment
Prehistoric Egypt (Prior to 3100 BC) Naqada III ("the protodynastic period"; approximately 3100–3000 BC) Early Dynastic Period (First–Second Dynasties) Old Kingdom (Third–Sixth Dynasties) First Intermediate Period (Seventh or Eighth–Eleventh Dynasties) Middle Kingdom (Twelfth–Thirteenth Dynasties) Second Intermediate Period (Fourteenth–Seventeenth Dynasties) New Kingdom (Eighteenth–Twentieth Dynasties) Third Intermediate Period (also known as the Libyan Period; Twenty-first–Twenty-fifth Dynasties) Late Period (Twenty-sixth–Thirty-first Dynasties) The Nile has been the lifeline for Egyptian culture since nomadic hunter-gatherers began living along it during the Pleistocene.	Període predinàstic (Abans de 3100 aC) Període Protodinàstic (aproximadament 3100-3000 aC) Període Arcaic d'Egipte (primera i segona dinasties) Antic Regne (dinasties 3-6) Primer Període Intermedi (Dinasties 7-11) Regne Mitjà (dinasties 12 i 13) Segon Període Intermedi (dinasties 14-17) Imperi Nou (divuitena-vintena dinasties) Tercer Període Intermedi (dinasties 21-25) (també conegut com el Període de Líbia) Període Tardà (dinasties 26-31) El Nil ha estat la línia de vida de la cultura egípcia des que els caçadors-recol·lectors nòmades van començar a viure a la vora del Nil durant el Plistocè.	682	Free translation[keep]
In the east, Roman–Persian Wars continued until 561 when Justinian's and Khusro's envoys agreed on a 50-year peace.	La guerra del 1650 amb la Confederació Iroquesa els va delmar moltíssim, de manera que el 1653 només restaven uns 500 attawendaroks a Pennsilvània i un miler de wendats i tionontatis.	450	Incorrect alignment
Jon runs him over out of revenge, but discovers that it was actually a cardboard cut-out before the real Mr. Pickles rips his face off and shoots him in the mouth.	Kane desperta aparentment il·lès, però durant un menjar abans d'entrar en hiperson comença a asfixiar-se convulsivament fins que un ésser emergeix violentament del seu pit, matant-lo i ocultant-se en la nau.	688	Incorrect alignment
This concept of "corrected lime potential" to define the degree of base saturation in soils became the basis for procedures now used in soil testing laboratories to determine the "lime requirement" of soils.	Aquest concepte de "potencial corregit de calcaris" per definir el grau de saturació de bases del sòl ha esdevingut la base del procediments de laboratori per determinar el requeriment de productes calcaris per encalcinar els sòls.	614	Valid translation
An even finer chronological distinction within Phase 3 is permitted by the settlement's architecture; the house type with underfloor channels, typical of Nevalı Çori strata I-IV, also characterises the "Intermediate Layer" at Çayönü, while the differing plan of the single building in stratum V, House 1, is more clearly connected to the buildings of the "Cellular Plan Layer" at Çayönü.	Dins de les primeres fases, una cronologia més afinada seria possible gràcies a l'arquitectura de l'assentament: l'habitatge tipus, amb canals subterranis, característica dels estrats I-IV de Nevali Çori, és igualment típica de l'«estrat intermedi» de Çayönü, mentre que la diferent planta de l'únic edifici de l'estrat V (casa 1) està més clarament connectada amb els edificis de "planta cel·lular" de Çayönü.	778	Free translation[keep]
She has been compared to Madonna, who has said that she sees herself reflected in Gaga.	Gaga ha estat comparada amb la cantant Madonna, que va admetre veure's reflectida en ella.	712	Free translation[keep]
Juan Fernández, nicknamed El Labrador, was a Spanish Baroque painter active between 1629 and 1636, specializing in still life painting.	Juan Fernández, anomenat el Labrador, va ser un pintor barroc espanyol actiu entre 1629 i 1636, especialitzat en la pintura de natures mortes.	814	Free translation[keep]
Underground tests in the United States continued until 1992 (its last nuclear test), the Soviet Union until 1990, the United Kingdom until 1991, and both China and France until 1996.	L'última detonació nuclear realitzada pels Estats Units (subterrània) tingué lloc a 1992, la Unió Soviètica continuà fins a 1990, el Regne Unit fins a 1991, i França i la Xina fins a 1996.	738	Free translation[discard]
Therefore, priorities are associated to each process.	Aquest tipus de sistemes assignen una prioritat a cada procés.	716	Valid translation
15% of respondents were over 65 years of age.	El 30% dels seus parlants té més de 65 anys.	576	Incorrect alignment
Nearly one third of Rossiya Bank's cash ($1 billion) was frozen in Cypriot accounts during this crisis.	Més de mil milions de hrívnies ($ 110 milions) pertanyents al pressupost regional de Crimea va ser congelats pel govern post-Euromaiden a Kíev.	414	Incorrect alignment
This is strange because, with a conventional shell and tube heat exchanger, there would be no risk of ash entering the turbine circuit.	Es tracta d'una mena de turbina que disposa de compressor i expansionador de gasos a base d'engranatges "sense fregament", evitant les pèrdues aerodinàmiques del àleps de les turbines convencionals.	524	Incorrect alignment

Some research suggests that women are more likely to volunteer for tasks that are less likely to help earn promotions,and that they are more likely to be asked to volunteer and more likely to say yes to such requests.	Alguns estudis realitzats suggereixen que a les dones no els agrada competir com als homes i potser per això tenen menys possibilitats de promoció.	260	Free translation[discard]
In August 1850, before Lind left England, sailing from Liverpool on the paddle steamer Atlantic, Barnum arranged for her to give two farewell concerts at the city's Philharmonic Hall.	A l'agost de 1850, abans que Lind abandonara Anglaterra, Barnum va organitzar per a ella dos concerts de comiat a Liverpool.	650	Free translation[discard]
The U.S. aircraft carrier, USS Saratoga, participated in the training exercises and the first two operations.	El portaavions americà USS Saratoga participà en els exercicis d'entrenament i a les dues primeres operacions.	714	Valid translation
Bituminite is not considered to be bitumen because its properties are different from most bitumens.	El xenó no és tòxic, però diversos dels seus compostos ho són altament a causa de les seues fortes propietats oxidants.	700	Incorrect alignment
The north and south elevations complement the west and east, with one centrally located continuous vertical window bay.	Es conserven, però, el mur perimetral septentrional i part del meridional, a més de l'absis, que presenta una interessant decoració llombarda.	462	Incorrect alignment
Kandahar Just average.	Tsaagan era dromaeosàurid de mida mitjana.	586	Incorrect alignment
Proceedings of the American Mathematical Society 17, 413–415 (1966)	Proceedings of the American Mathematical Society, 17, 2, 01-02-1966, pàg. 413–415.	788	Free translation[discard]
Tokyo Metropolitan Police Department Kidotai (Riot Police) water cannon unit.	Departament de policia metropolitana de Tòquio Kidotai (policia antidisturbios) unitat de canó d'aigua.	716	Translation error
As well as editing feature films such as Malvaloca, Pallejá was noted for his work on documentaries.	No obstant això, seguí present a les festes asturianes tal com documenta Jovellanos en els seus escrits.	380	Incorrect alignment
There was no commentary in the game.	No hi ha veus en el joc.	424	Free translation[keep]
XpatLoop.	Trèvol.	0	Incorrect alignment
For leisure, he would climb a nearby hill, Roseberry Topping, enjoying the opportunity for solitude.	En els moments d'oci pujava dalt d'un turó proper, Roseberry Topping, on podia gaudir de la solitud.	552	Free translation[keep]
The whole magazine was typeset in Helvetica.	La redacció de la revista estava situada a les Galeries Laietanes.	424	Incorrect alignment
The initiative was taken by the Royal Society (United Kingdom) which resulted in a meeting in London in June 1986 of Arnold Burgen (United Kingdom), Hubert Curien (France), Umberto Columbo (Italy), David Magnusson (Sweden), Eugen Seibold (Germany) and Ruud van Lieshout (the Netherlands) – who agreed to the need for a new body.	La Royal Society britànica va organitzar després una reunió a Londres al juny de 1986 a la qual van assistir Arnold Burgen (Regne Unit), Hubert Curien (França), Umberto Colombo (Itàlia), David Magnusson (Suècia), Eugen Seibold (Alemanya) i Ruud van Lieshout (Països Baixos), que van coincidir en la necessitat d'una Acadèmia Europea.	804	Free translation[discard]
Formed after the dissolution of Homme's previous band, Kyuss, Queens of the Stone Age developed a style of riff-oriented, heavy rock music.	La banda nasqué després de la separació de Kyuss, l'anterior banda de Homme, i adoptà un estil musical dur i orientat als riffs.	588	Free translation[discard]
Then, in 2001, the ASMS became the Swiss Cetacean Society (SCS).	Aquell mateix any l'ACEC va passar a ser el Centre Excursionista de Catalunya (CEC).	314	Incorrect alignment
An art of numerical analysis is to find a stable algorithm for solving a well-posed mathematical problem.	Un art d'anàlisi numèrica és trobar un algorisme estable per resoldre un problema matemàtic ben condicionat.	732	Valid translation
With the Army Medical School reopened in 1901, he became an assistant instructor for Military Hygiene.	Després de la creació de l'Acadèmia General Militar en 1927, l'Acadèmia d'Infanteria es converteix en Acadèmia d'Aplicació d'Infanteria.	402	Incorrect alignment
There are 6,415,851,530,241 topologically distinct convex heptadecahedra, excluding mirror images, having at least 11 vertices.	Hi ha 6.415.851.530.241 heptadecàedres convexos topològicament diferents, excloent les imatges de mirall, que tenen almenys 11 vèrtexs.	770	Valid translation
Gagik (Georgian: გაგიკი) (died 1058) was a King of Kakheti and Hereti in eastern Georgia from 1039 to 1058.	Gagik de Kakhètia (en georgià : გაგიკი) fou un rei de Kakhètia del 1039 a 1058.	0	Free translation[discard]

Jackass 3.5 was compiled from outtakes shot during the making of the third film.	Jackass 2.5 és una "tercera pel·lícula" sorgida per escenes rebutjades de la segona pel·lícula.	422	Incorrect alignment
The members of the chapter, which today represents in part the whole Church of Ireland, hold one of four dignities or special offices, or one of 24 prebends (22 regular, 2 ecumenical as noted below).	En l'actualitat 46 publicacions d'arreu de l'illa de Mallorca formen part de l'APFM, de les quals set són de periodicitat setmanal, quatre són quinzenaris, 22 d'elles són mensuals, 7 bimestrals i 7 trimestrals.	554	Incorrect alignment
This is a network of providers for care that is not of an emergency nature.	Sòl de naturalesa rústica és aquell que no sigui de naturalesa urbana.	760	Incorrect alignment
The temple probably acted as a landmark for navigation of their ships.	Probablement hi havia un temple de Venus per a donar bona sort als mariners.	764	Incorrect alignment
The program was soon cancelled due to political turmoil.	El programa va ser poc després cancel·lat a causa de l'agitació política.	280	Valid translation
Now a millionaire, Louisa vows never to marry again.	Torna ric, i el pare ja no s'oposa al matrimoni.	548	Incorrect alignment
Available in four grades: Type 440A—has the least amount of carbon making this the most stain-resistant.	En segon lloc, és un estel de mercuri-manganès; amb una temperatura efectiva de 14.400 K, és considerada freqüentment l'estel més calent dins d'aquesta classe.	314	Incorrect alignment
He sold his share of Simtek in 1991, when elected president of the FISA.	Va vendre la seva porció de Simtek l'any 1991 moment en què es convocaven unes eleccions a la presidencia de la FISA.	552	Free translation[discard] [discard]
He returned to Ashby where he practiced as a physician for some years, kept a school and studied astronomy.	Va retornar a Ashby on va practicar la Medicina durant quatre anys, va dirigir una escola i va estudiar astronomia pel seu compte.	812	Free translation[discard]
Tammany boss William M. Tweed was an American politician who ran what is considered now to have been one of the most corrupt political machines in the country's history.	La Societat de Sant Tammany, dirigida per William Tweed, va liderar la política d'una manera que avui dia es consideraria corrupta.	694	Free translation[discard]
Corbicula Megerle von Mühlfeld, 1811 (central and southern Africa, Central and southern Asia) Geloina (southern Asia with Malaysia) Cyrenodonax (southern China, Vietnam) Cyrenobatissa (northern Vietnam) Batissa (Malaysia and Indonesia) Corbiculina (eastern Australia) Solielletia (Ethiopia) Polymesoda Rafinesque, 1820 (Gulf coast, and Atlantic coast of northern South America) Neocorbicula Fischer, 1887 (Gulf coast, and Atlantic coast of northern South America) Pseudocyrena (Caribbean side of Central America) Egetaria (Atlantic coast of South America) Villorita (eastern parts of South America) Glaubrecht, M., et al. (2007).	Corbicula Megerle von Mühlfeld, 1811 (centre i sud d'Àfrica i d'Àsia) Geloina (sud d'Àsia amb Malàisia) Cyrenodonax (sud de la Xina, Vietnam) Cyrenobatissa (northern Vietnam) Batissa (Malàisia, Indonèsia) Corbiculina (est d'Austràlia) Solielletia (Etiòpia) Polymesoda Rafinesque, 1820 (Golf, costes atlàntiques del nord de Sud-amèrica) Neocorbicula Fischer, 1887 (Golf, costes atlàntiques del nord de Sud-amèrica) Pseudocyrena (costat Carib de Centreamèrica) Egetaria (costes atlàntiques de Sud-amèrica) Villorita (est de Sud-amèrica)	758	Free translation[keep]
Geologically Galdhøpiggen, like most of Southern Norway's mountain ranges, belongs to the Caledonian folding.	El Galdhøpiggen pertany geològicament, com la majoria de les serralades del sud de Noruega, a l'orogènesi caledoniana.	798	Valid translation
The PKT is usually fed from 250-round ammunition boxes.	Una caixa de munició amb cintes de 250 cartutxos.	574	Free translation[discard]
The first State constitution was created in 1776, shortly after the country declared its independence from Great Britain.	Aquest dia commemora la signatura de la Declaració d'Independència el 1776 en la qual el país va proclamar la seva separació formal de l'Imperi britànic.	590	Free translation[discard]
Given that the MMX's 64-bit MMn registers are aliased to the FPU stack and each of the floating point registers are 80 bits wide, the upper 16 bits of the floating point registers are unused in MMX.	També perquè els registres MMX de 64bits MMn són àlies de pila de la FPU, i cada un dels registres de la pila té una amplada de 80 bits, els 16 bits registres superiors de la pila no utilitzants en MMX, i aquests bits són establerts tots a uns, que el fan semblar com un NaN o infinits des del punt de vista de coma flotant.	0	MT translation
The political situation became critical: the workers were protesting against the government, and Ady saw a revolution approaching.	Altrament, la situació política es complicà amb severitat; els obrers endegaren un seguit de protestes contra el govern del moment, fet que permeté Ady d'albirar i ensumar l'esclat d'una revolució propera.	752	Free translation[discard]
He was found by Superboy, who suspected the youth may be his older brother.	Fou probablement fill d'Antef VI si bé alguns pensen que podia ser el seu germà.	506	Incorrect alignment

Genetic studies indicate the genus Leopardus forms a distinct clade within the feline subfamily, and first evolved in South America around 8 million years ago.	Els estudis genètics indiquen que el gènere Leopardus forma un clade propi dins de la subfamília dels felins, i que evolucionà inicialment a Sud-amèrica fa entre 10 i 12 milions d'anys.	628	Translation error
Carboxylic acids can also be obtained by the hydrolysis of nitriles, esters, or amides, in general with acid- or base-catalysis.	Els àcids carboxílics també poden obtenir-se per la hidròlisi dels nitrils, èsters, o amides, generalment amb catàlisi àcida o bàsica.	772	Valid translation
Theódór Elmar "Teddy" Bjarnason (born 4 March 1987 in Reykjavík, Iceland) is an Icelandic footballer who currently plays as a central midfielder for Gazişehir Gaziantep.	Theodor Elmar "Teddy" Bjarnason (Reykjavík, Islàndia el 4 de març de 1987) és un futbolista islandès que juga com a migcampista esquerre amb el AGF de Dinamarca.	810	Free translation[discard]
Conrad considered this to be the only way to end the long-standing feud between Saxons and Franks and to prevent the dissolution of the Empire into smaller states based on the German stem duchies.	Conrad considerava aquesta l'única manera per posar fi a la llarga disputa entre els francs i els saxons i prevenir la dissolució de l'imperi en estats més petits creats sobre la base dels ducats tribals alemanys.	742	Valid translation
In an attempt to molify his critics, Obiang announced a new cabinet, giving minor portfolios to some people identified by the government as opposition figures.	En un intent de fer callar als seus crítics, Obiang va anunciar el seu nou gabinet, donant carteres de menor importància a algunes persones identificades pel govern com de l'oposició.	874	Valid translation
Book of Armagh.	The Book of Armagh (anglès)	720	Free translation[discard]
He then lost Mistilteinn in the water out of witchcraft.	Va morir al continent en la disputa amb el sàtrapa persa Oroetes.	596	Incorrect alignment
On 20 July 1917, the signing of the Corfu Declaration foreshadowed the unification of Montenegro with Serbia.	El 20 de juliol de 1917, la Declaració de Corfú va ser signada, va declarar la unificació de Montenegro amb el Regne de Sèrbia.	682	Free translation[discard]
He said it is time to get started in theatre as a professional.	Sempre diu que començà en el teatre regional professionalment.	398	Free translation[discard]
There have been few health surveys of the individual Belgian Shepherd varieties.	No existeixen malalties típiques en aquesta varietat de pastor belga.	480	Free translation[discard]
Les noces improvisées (libretto by Armand Liorat and Albert Fonteny); premiered at the Théâtre des Bouffes-Parisiens, Paris, 13 February 1886.	Les noces improvisées, llibret d'Armand Liorat i Albert Fonteny, Théâtre des Bouffes-Parisiens, París, 13 febrer 1886.	804	Free translation[discard]
The Mayans engineered some of the most important monuments in Mesoamerica.	L'albufera inclou alguns dels monuments megalítics més importants dels conservats a Menorca.	692	Incorrect alignment
Most ancient cultures used various methods of divination to attempt to circumvent randomness and fate.	Moltes cultures antigues empraven diversos mètodes d'endevinació per intentar esquivar l'atzar i el destí.	832	Valid translation
He was elected as member of the National Assembly of Pakistan for the first time in Pakistani general elections, 1985 as an independent candidate.	Fou elegit diputat dins de les files del Partit Conservador per primer cop a les eleccions generals espanyoles de 1886, amb només 25 anys.	460	Incorrect alignment
There is also a second-floor side entrance accessed from an outside stairway.	També presenta una terrassa davantera, a la qual s'accedeix per unes escales directament des de l'exterior.	424	Free translation[discard]
You ceaselessly move and strike, and are always trying to get to the outside by turning.	Han de pujar i baixar sense parar i cercar sempre el suport per als seus companys.	502	Incorrect alignment
Some tendencies that define modern Western societies are the existence of political pluralism, laicism, generalization of middle class, prominent subcultures or countercultures (such as New Age movements), increasing cultural syncretism resulting from globalization and human migration.	Algunes tendències que defineixen les societats occidentals modernes són l'existència de pluralisme polític, subcultures o contracultures prominents (com els moviments nova era) i l'augment del sincretisme cultural fruit de la globalització i migració humana.	876	Valid translation
El Panteón de Marinos Ilustres.	Els pirates dels mars imaginaris.	0	Incorrect alignment
These have corrugated iron roofs with small roof lanterns and regularly spaced sash windows.	Conserva finestres amb tancament de fusta tornejada tradicionalment i petites reixes de ferro forjat.	674	Incorrect alignment
One major focus of enquiry is the prediction of eruptions; there is currently no accurate way to do this, but predicting eruptions, like predicting earthquakes, could save many lives.	Una via d'investigació majoritària és la predicció de les erupcions; actualment, no hi ha manera de realitzar aquestes prediccions, però preveure els volcans, igual que preveure els terratrèmols, pot arribar a salvar moltes vides.	894	Free translation[keep]

The team was officially renamed Footwork in 1991, and secured a deal to race with Porsche engines, but the car was woefully noncompetitive and in 1992 they switched to a Ford V8, and then to Mugen engines.	L'equip va prendre el nom oficial de Footwork el 1991, i va obtenir un contracte per a córrer amb els motors de Porsche, amb resultats desastrosos, i el 1992 van passar a Mugen.	782	Free translation[discard]
They were housed in four brick barracks.	Allí fou trobat per uns barquers.	572	Incorrect alignment
3 go on trial over slaying at embassy.	Es celebrà el judici al Jutjat d'Instrucció número 1 de Palma.	400	Incorrect alignment
Radio-carbon analysis returned a date of c.8300, approximately 1000 years older than the northern site, making it the earliest known dwelling in Scotland.	L'habitatge, basat al voltant d'una massa ovalada d'aproximadament 7 metres (23 ft) de longitud, ha estat datada al voltant del 8300 aC, convertint-la en l'habitatge més antic conegut a Escòcia.	568	Free translation[discard]
Pere Gimbernat i Quintana was at an age where he could not take too much risk and he didn't have the funds to buy his uncles part of the business, the situation became untenable all ending in the closure of the factory in Figueres.	Pere Gimbernat i Quintana es trobava en una edat en què no podia assumir gaire riscos i tampoc disposava dels recursos econòmics per a comprar la part dels seus oncles, fent-se aquesta situació insostenible i desembocant tot plegat en el tancament de la fàbrica de Figueres.	788	Valid translation
The bark is dark brown to blackish, becoming fissured in older trees, and frequently forking near the base.	La pupa és lliure, i s'acostuma a trobar a terra entre la fullaraca, molt pròxima als arbres infestats.	812	Incorrect alignment
Kohonen valittiin maailman parhaaksi salibandyn pelaajaksi.	«Coentrão va ser designat millor jugador del partit.» (en castellà).	0	Wrong language id.
Abdominal ultrasound from MedlinePlus.	Incontinència urinària en MedlinePlus.	610	Translation error
Aivars Lembergs was the candidate of the Union of Greens and Farmers for the position of Prime Minister in 2006, before being charged with corruption, fraud, bribery, money laundering and abuse of elected office on 20 July 2006.	Lembergs va ser el candidat de la Unió de Verds i Agricultors per al càrrec de primer ministre el 2006, abans de ser acusat de corrupció, frau, suborn, blanqueig de diners i l'abús de poder el 20 de juliol de 2006.	704	Free translation[keep]
In addition, critics contend that textbooks written in this manner are perceived by students as dry and uninteresting and discourage students from reading history, creating motivational barriers to learning.	És interessant remarcar també que en el seu escrit sobre la restauració estableix una profunda crítica a aquells intel·lectuals que estigmatitzaven i no permetien l'estudi de l'Edad Mitjana.	810	Incorrect alignment
Freddie Prinze Jr. Twitter.	El repartiment va incloure: Freddie Prinze Jr.	0	Translation error
Dawson was undoubtedly indebted to and supportive of Edward but the account of his close colleague William Evans appears to clear him of any unethical manipulation in the matter.	Això fou clarament un precursor de la seva pròpia fabricació d'aigua de soda i, probablement, conduí als estudis quantitatius del seu fill William Henry sobre la dissolució de gasos per fluids que culminaren amb la llei de Henry.	338	Incorrect alignment
This nebula was first discovered in 2010 by Mubdi Rahman and Norman Murray from the University of Toronto.	Narrative Science és un projecte que va arrencar durant el mes de gener de 2010, creat per Stuart Frankel amb la col·laboració de Kris Hammond i Larry Birnbaum, professors de la Northwestern University.	430	Incorrect alignment
University of Mannheim School of Social Sciences.	Facultat de Medicina de la Universitat de Mannheim.	658	Translation error
He started working for his father's construction company.	Va començar a treballar a l'empresa de construcció de son oncle.	656	Translation error
After the 2008 Soacha discoveries, defense minister Juan Manuel Santos denied knowledge of the scheme, fired 27 officers including three generals and changed the army's body count system.	Després de la descoberta de Soacha de 2008, el ministre de defensa Juan Manuel Santos va negar el seu coneixement de la situació i va acomiadar 27 agents, incloent tres generals i va canviar el sistema de recompte del cos de l'exèrcit.	762	Valid translation
Graf also requested Fritz Walter, who later captained the West German World Cup team of 1954.	Entre els seus jugadors estava Fritz Walter, futur capità de l'equip alemany guanyador de la Copa del Món de 1954.	730	Free translation[discard]
The Hungarian minority of Romania (Hungarian: Magyarok Romániában, Romanian: Maghiarii din România) is the largest ethnic minority in Romania, consisting of 1,227,623 people and making up 6.1% of the total population, according to the 2011 census.	Els hongaresos de Romania són la principal minoria ètnica de Romania, consistent en 1.431,807 persones que formen el 6,6% del total de la població, segons el cens del 2002.	560	Free translation[discard]
In 1914, he was transferred back to the United States, where he was stationed in Eagle Pass, Texas and also took a part in Pancho Villa Expedition under the command of	El 1914 tornà als Estats Units, sent destinat a Eagle Pass, Texas, participant en l'Expedició de Pancho Villa, sota el comandament del general John J. Pershing.	678	Translation error

Brigadier General John Joseph Pershing (later destined to achieve the highest rank in the United States Army, that of General of the Armies).			
It was mentioned in the Jewish Tosefta (Demai 1) as being included in the boundary of the southern mountains of Judea.	Fou esmentat al Tosefta jueu (Demai 1) com a inclòs al límit de les muntanyes del sud de Judea.	546	Valid translation
The King spent several days there, and was entertained by the Cardinals.	Donà diverses vegades la volta al món i fou aplaudit pels públics dels quatre punts cardinals.	740	Incorrect alignment
The United States Department of Defense reported that in Taiwan, "proponents of strikes against the mainland apparently hope that merely presenting credible threats to China's urban population or high-value targets, such as the Three Gorges Dam, will deter Chinese military coercion."	En concret, va indicar que "els proponents d'atacs contra el continent semblen esperar que la mera presentació d'amenaces creïbles contra la població urbana de la Xina continental o objectius d'alt valor, com la presa de les Tres Gorges, dissuadiria la coacció militar xinesa."	788	Free translation[discard]
For the occasion, Bernstein reworded Friedrich Schiller's text of the Ode to Joy, substituting the word Freiheit (freedom) for Freude (joy).	Per a l'ocasió, Bernstein va fer un canvi en l'Oda a l'Alegria de Friedrich Schiller, substituint la paraula "alegria" (Freude) per "llibertat" (Freiheit).	876	Free translation[discard]
The match could be divided into halves, the first won convincingly by Fischer, and the second a close battle.	El matx es pot dividir en dues meitats, la primera la va guanyar de manera convincent en Fischer, i la segona, en canvi, fou una renyida batalla.	630	Free translation[keep]
Moffett said the slowing rate indicated that online sources were not making people drop cable as quickly.	Moffett va dir que l'índex alentit indicava que les fonts online no feien que els consumidors deixessin el cable tan ràpidament.	774	Valid translation
In and around June 2010 a "coup" was related to website visitors (in accordance with a spurious interpretation of Romans 13) whereby Pastor Harry Harkwell took over leadership.	Al voltant del juny de 2010 es va produir un cop d'estat, relatat pels visitants de la pàgina web i seguint alguns passatges de la Bíblia, en el qual el pastor Harry Harkwell va prendre el lideratge.	566	Free translation[discard]
Yearly, he gave away thousands of pounds, much of it to clergymen to distribute in their parishes.	Anualment donava milers de lliures esterlines, la majoria per sacerdots que les distribuïen entre els seus parroquians.	642	Valid translation
Kari'nja has a typical 6 vowel system after *ô merged with *o, being a e i o u ï.	El kali'na té un sistema típic de 6 vocals després que *ô es fusiona amb *o, essent a e i o u ï.	0	Valid translation
The film was directed by Kevin Macdonald from a screenplay written by Matthew Michael Carnahan, Tony Gilroy, Peter Morgan, and Billy Ray.	Aquesta última la va dirigir el Kevin Macdonald i el guió va ser escrit pel Matthew Michael Carnahan, el Tony Gilroy, el Peter Morgan i el Billy Ray.	850	Free translation[keep]
France halts delivery of the first of two Mistral-class amphibious assault ships to the Russian Navy due to circumstances in Ukraine.	Així mateix, ha transcendit la compra dels vaixells d'assalt amfibi que França havia construït per Rússia de la classe Mistral.	462	Free translation[discard]
The manner of Siger's death, which occurred at Orvieto, is not known.	La notícia de la mort del sultà fou coneguda pels croats a Damietta, no se sap com.	402	Incorrect alignment
Falcon 1 achieved orbit on its fourth attempt, in September 2008 with a mass simulator as a payload.	El Falcon 1 arribà a l'òrbita en el seu quart intent el dia 28 de setembre del 2008, amb un simulador de massa com a càrrega.	574	Free translation[keep]
On 20 December, a three-man ETA commando unit disguised as electricians detonated the explosives by command wire as Blanco's Dodge Dart passed.	El 20 de desembre de 1973, un comando d'ETA de 3 homes dissimulava mentre els electricistes feien detonar els explosius quan el Dodge Dart de Carrero Blanco passava.	802	Free translation[discard]
This led to roles in bigger productions such as the action film Lara Croft: Tomb Raider (2001), the crime thriller Road to Perdition (2002), the crime thriller Layer Cake (2004), and the Steven Spielberg historical drama Munich (2005).	Va debutar al cinema amb The Power of One (1992) i anys més tard va atreure l'atenció de la indústria cinematogràfica internacional en intervenir en pel·lícules com Lara Croft: Tomb Raider (2001), Road to Perdition (2002) de Sam Mendes, Crim organitzat (Layer Cake) (2004), Munic (2005) de Steven Spielberg, Resistència (2008) o Millennium: Els homes que no estimaven les dones (2011).	774	Free translation[discard]
In retirement, he returned to work as a saddler.	De retorn a la vida civil, tornà a Iakut per treballar en un sovkhoz.	476	Incorrect alignment
The drama of the time also followed the same stylistic evolution as poetry and prose — expressionism, followed by a return to realistic, civilian theater (František Langer, Karel Čapek).	El teatre de l'època també va seguir la mateixa evolució estilística que la poesia i la prosa —expressionisme, seguit d'una volta al teatre realista i civil, com per exemple en František Langer i Karel Čapek.	854	Valid translation
‹ The template Infobox NRHP is being considered for merging. › Hampstead, also known as Henry Rose House, is a historic home located at Jerusalem in Yates County, New York.	Hampstead, també coneguda com a Henry Rose House, és una casa històrica ubicada a Jerusalem al Comtat Yates (Nova York).	660	Wrong tokenization

PepsiCo	Pequinès	0	Incorrect alignment
Blues performers explored a range of musical genres, as can be seen, for example, from the broad array of nominees of the yearly Blues Music Awards, previously named W.C. Handy Awards or of the Grammy Awards for Best Contemporary and Traditional Blues Album.	En els anys noranta, els intèrprets de blues van explorar un ampli rang de gèneres musicals, com es pot veure en les nominacions per als premis anuals Blues Music Awards, anomenats amb anterioritat W. C. Handy Awards, o en les nominacions per als premis Grammy en les categories de Millor disc de blues contemporani i Millor disc de blues tradicional.	596	Free translation[discard]
A few days after arriving at Nöteborg, Peter sent a detachment of 400 men of the Preobrazhensky Regiment to take up positions closer to the fort and prepare for the arrival of a greater force.	Uns quants dies després d'arribar a Nöteborg, Pere envià un destacament de 400 homes del Regiment de Preobrazhensky per ocupar posicions més properes al fort i preparar-se per l'arribada del contingent principal.	864	Valid translation
The only other complete Han dynasty sundial, in the collection of the Royal Ontario Museum, also has a Liubo pattern carved on it.	Un altre rellotge de sol de la dinastia Han, únic a la col·lecció del Museu Reial d'Ontari, també té un disseny liubo gravat en ell.	554	Valid translation
In 2011, it had a student population of 14,000, with a branch in Tel Aviv.	En 2011, el campus tenia una població estudiantil de 14.000, amb una sucursal a Tel Aviv.	834	Valid translation
In her class of forty-five students, Parlow was the only female.	La seva classe constava de quaranta-cinc estudiants on Parlow era l'única noia.	570	Free translation[keep]
A small number assembled nonetheless, and, in accordance with their wishes, Gregory again excommunicated Henry.	No obstant això, es reuní un petit nombre i d'acord amb els seus desitjos, Gregori excomunicà de nou Enric.	666	Valid translation
If we do want it to succeed, we should surely put behind it the full force of our influence.	Emocioneu-vos, perquè necessitarem tot el nostre entusiasme.	184	Incorrect alignment
Bennett secretly hired Arcand as his chief electoral organizer in Quebec for the 1935 federal election.	Fins i tot després d'iniciar aquesta campanya política, Bennett va nomenar Arcand organitzador de les eleccions al Quebec en secret a temps per a les eleccions federals de 1935.	688	Free translation[discard]
The second KKK was a formal fraternal organization, with a national and state structure.	El segon KKK fou una organització més formal, que comptava amb membres registrats i amb una estructura estatal i nacional.	760	Free translation[keep]
Election silences are observed in the following countries, amongst others.	Existeixen jornades de reflexió en els següents països, entre altres.	550	Valid translation
Atilius, a freedman, built an amphitheatre at Fidenae in the reign of Tiberius, which collapsed, killing between twenty and fifty thousand spectators.	Atilius, un llibert que va construir un amfiteatre a Fidenes en temps de l'emperaor Tiberi, el qual es va esfondrar causant la mort a uns 20.000 o 50.000 espectadors i va ser exiliat com a càstig.	334	Free translation[discard]
The vast majority of the land in the watershed is rangeland, but there is also some cropland, pastureland and developed areas.	La major part de les terres productives són conreus de secà, però també n'hi ha alguns de regadiu i algunes pastures.	732	Valid translation
The Norwegian name Tromsdalstinden means "the peak above Tromsdalen", while the Sámi name is made up from the components Sálaš and Oaivi.	El nom noruec Tromsdalstinden significa "el pic anterior de Tromsdalen", mentre que el nom sami Sálašoaivi es compon dels components Salas i Oaivi.	738	Valid translation
The World Reference Base for Soil Resources (WRB) aims to establish an international reference base for soil classification.	La Base de Referència Mundial pels Recursos del Sòl (en anglès: World Reference Base for Soil Resources (WRB)), és el sistema taxonòmic estàndard internacional de la classificació dels sòls.	792	Free translation[discard]
He was named president of the Association of Banks of Northeastern Spain and began contributing regularly to the publication España Bancaria.	Va ser vicesecretari de l'Associació de Banquers de Barcelona- Asociación de Bancos del Nordeste de España i va col·laborar amb la publicació Espanya Bancària.	720	Free translation[discard]
More precisely: for bravery and selflessness displayed during combat operations aimed at the arrest of violators of the State Border of the USSR; for the leadership of border protection units while ensuring the inviolability of the borders of the USSR; for a high degree of vigilance and proactive actions which resulted in the arrest of violators of the State Border of the USSR; for the skilful organization of border service units and exemplary work to strengthen the borders of the USSR; for the excellent performance of military duties associated with the protection of the state borders of the USSR; for active	Era atorgada als guàrdies fronterers, homes de servei de les Forces Armades i civils per la seva actuació i serveis rendits en la defensa de les fronteres de la U.R.S.S: Per la valentia i l'abnegació en les operacions militars en la detenció dels infractors de les fronteres de l'URSS Per la direcció sàvia de les operacions militars a la frontera, juntament amb la defensa de la inviolabilitat de la frontera de l'URSS Per l'alta vigilància i les accions iniciadores que han tingut com a resultat la detenció dels infractors de les fronteres de l'Estat Per l'organització hàbil del servei fronterer i el treball de reforç de les fronteres de	606	Free translation[discard]

assistance to border protection forces in their combat assignments aimed at the protection of the state borders of the USSR.	l'URSS Pels serveis irreprotxables en la vigilància de les fronteres de l'URSS, per l'ajut actiu a les tropes de la defensa de fronteres en la seva activitat de combat per la vigilància de les fronteres de l'Estat La primera concessió va ser el 22 d'agost de 1950 al tinent D.V. Ignatev.		
The system flourished throughout the French and Spanish colonial periods, reaching its zenith during the latter, between 1769 and 1803.	El sistema va prosperar durant els períodes colonials francès i espanyol, i pel que sembla va arribar al zenit durant aquest últim, entre 1769 i 1803.	736	Free translation[keep]
Funky Fries and other foods that flopped.	«Freixe i els seus derivats Freixenet i altres».	540	Incorrect alignment
In early November, Chatham opened fire at long range and set fire to Somali, but she failed to hit Königsberg, which promptly moved further upstream.	A principis de novembre, el HMS Chatham va obrir foc a llarga distància contra els vaixells i va incendiar al vaixell somalí, però no va encertar al SMS Königsberg, que es va moure ràpidament.	742	Free translation[discard]
It is normal practice to incorporate both regenerative and rheostatic braking in electrified systems.	És pràctica habitual utilitzar conjuntament el fre regeneratiu i el reostàtic.	702	Free translation[discard]
In the 1966 novel The Last Battle, Cornelius Ryan records that Blanter accompanied the Red Army into Berlin during the last days of the war and the collapse of Nazi power.	En la novel·la "L'última batalla" de 1996, Cornelius Ryan registra que Blànter va acompanyar l'Exèrcit Roig a Berlín durant els últims dies de la guerra i el col·lapse del poder nazi.	734	Valid translation
This dog is good with, and protective of, children with whom they were brought up.	Manso i pacient amb els nens, als quals vol i protegeix.	562	Free translation[discard]
Stacy ends up winning the competition.	Al final, l'Stacey guanya la competició.	564	Free translation[keep]
The Flavr Savr failed to achieve commercial success and was withdrawn from the market in 1997.	El model de Philips no va aconseguir l'èxit que s'esperava i va desaparèixer del mercat el 1996.	504	Incorrect alignment
Internet History Project biography, 2003.	«Internet History Project biography, 2003» (en anglès), 09-10-2003. .	722	Free translation[discard]
In dynamic analysis, static reduction refers to reducing the number of degrees of freedom.	L'aparició d'articulacions redueixen el grau de hiperestaticitat ampliant el nombre de graus de llibertat.	588	Valid translation
The Mobile Remote Servicer Base System (MBS) is a base platform for the robotic arms.	El Sistema base mòbil (MBS, Mobile Base System en anglès) és una plataforma base per al braç robòtic.	850	Free translation[discard]