Prevalencianització de l'eixida de traductors automàtics al català

Nom: Héctor Navarro Isaac

Línia d'investigació: Tecnologies de la Traducció.

Tutor: Mikel Lorenzo Forcada Zubizarreta

Data: 06/06/2022

Treball de Fi de Grau de Traducció i Interpretació

Prevalencianització de l'eixida de traductors automàtics al català

Héctor Navarro Isaac

hni1@alu.ua.es

RESUM

Els traductors automàtics s'utilitzen cada vegada més, però si els volem emprar per a traduir al català, els textos resultants solen presentar formes i expressions pròpies de la varietat oriental, cosa que fa la postedició més pesada si necessitem el text en algun dels estàndards valencians. Aquest treball explora la possibilitat de "prevalencianitzar" l'eixida d'aquests traductors mitjançant operacions de substitució basades en expressions regulars per tal d'alleugerir la postedició. Amb aquest objectiu s'ha recopilat un corpus d'entrenament de textos periodístics que ha estat traduït al català automàticament i a partir del qual s'ha elaborat una llista de mots a substituir basada en la freqüència. A continuació s'han implementat les operacions de substitució en l'entorn de les macroinstruccions del processador de text LibreOffice i finalment han estat provades en quantitats suficients de text. A partir dels resultats es pot concloure que la macroinstrucció realitza la seua funció correctament, encara que presenta algunes limitacions.

ABSTRACT

"Pre-valencianising the output of machine translators into Catalan"

Machine translation is being increasingly used lately, but if we want to use it to translate into Catalan, the resulting texts tend to have forms and expressions typical of the Eastern variety, which makes post-editing more laborious if we need the text in

one of the Valencian standards. This paper explores the possibility of pre-valencianising the output of these translators by means of substitution operations based on regular expressions to lighten the post-editing process. To this end, a training corpus of journalistic texts has been compiled and translated into Catalan automatically, from which a list of words to be substituted based on frequency has been elaborated. The substitution operations were then implemented in the LibreOffice word processor macro instructions environment and finally tested on sufficient quantities of text. It can be concluded from the results that the macro-operation performs its function correctly, although it has some limitations.

Paraules clau: Traducció automàtica, català, valencià, expressions regulars, postedició,
Keywords: Machine translation, Catalan, Valencian, regular expressions, postediting

Índex

1. Introducció

Els traductors automàtics s'han convertit en una part cada vegada més important del món actual, fins al punt que, qui més qui menys, tots els hem utilitzat alguna vegada.

I, encara que els avantatges dels traductors automàtics són innegables, si volem fer-ne ús d'aquests per traduir al valencià, trobem un problema: l'eixida de molts sistemes de traducció basats en corpus (com ara Google Translate o el traductor neuronal de Softcatalà) sol contenir formes i estructures pròpies dels estàndards escrits de Catalunya, ja que els textos paral·lels que s'usen per a entrenar-los provenen majoritàriament d'aquesta àrea. Per tant, si volem traduir a algun dels estàndards emprats al País Valencià, la postedició és més pesada perquè comporta la substitució de les formes que són diferents en un estàndard i l'altre.

L'objectiu d'aquest projecte és explorar la possibilitat de la "prevalencianització" automàtica de l'eixida d'aquests sistemes per alleugerir la postedició, procés que es realitzarà mitjançant operacions de substitució basades en expressions regulars per fer servir-les tant sobre textos plans com en entorns de processament de textos.

Per tal d'aconseguir això s'ha procedit, en primer lloc, a estudiar les operacions de postedició més comunes necessàries per a produir una traducció en valencià a partir de l'eixida de Google Translate i classificant-les en una taxonomia. Per a fer això s'ha elaborat un corpus d'entrenament de 100000 paraules[1] compost per textos periodístics. A continuació s'ha construït un catàleg d'operacions de substitució basades en expressions regulars que cobrisquen la major part de les operacions de l'apartat anterior i ha estat implementat usant macroinstruccions en l'entorn del processador de textos LibreOffice. Finalment, s'ha fet una avaluació quantitativa de l'efecte corrector d'aquest procediment automàtic de

[1] Els corpus estan recopilats a una carpeta compartida a l'annex

prevalencianització sobre un altre corpus també compost per textos periodístics, aquesta vegada de 10000 paraules.

1.1 Breu definició de les expressions regulars

Abans de començar, és important entendre què són les expressions regulars. Les expressions regulars són patrons que descriuen una certa quantitat de text i el seu nom prové de la teoria matemàtica en la qual estan basades (Goyvaerts). Les expressions regulars s'utilitzen per cercar i extraure o manipular textos basant-se en patrons regulars, que en aquest cas esdevenen patrons lingüístics. La seua sintaxi pot ser simple o molt complexa, depenent dels patrons (Guzmán, 2007).

Dit d'una altra manera, les expressions regulars són eines que ens permeten, a partir d'unes instruccions que nosaltres introduïm, trobar combinacions de caràcters dins d'una cadena de text de qualsevol dimensió. Per tant, podem cercar mots sencers, combinacions d'aquests, o qualsevol patró que puguem imaginar. Açò ens pot servir, per exemple, des d'identificar totes les paraules d'un text fins a trobar formes verbals específiques com poden ser les del subjuntiu acabades en -i, com s'ha fet en el present treball per tal de substituir-les.

1.2 Estàndards lingüístics al País Valencià

Una de les peculiaritats del present treball té a veure amb la situació lingüística del País Valencià, i és que existeix una varietat d'estàndards. Per culpa de diverses motivacions històriques i polítiques, els debats amb relació als models formals al territori valencià tenen un origen més ideològic que filològic (Mas, 2010). És per aquesta raó que trobem taxonomies aproximades com la de Mas (2008) que distingeix "quatre models, basats lingüísticament en un percentatge de més a menys formes pròpies del valencià." Aquests models són:

- Secessionista: es distingeix de la resta perquè és l'únic que queda totalment fora de la codificació fabriana. D'aquesta manera fa una proposta no ja d'estàndard sinó de normativa, amb l'objectiu explícit de distingir-se al màxim de la consolidada.
- Particularista: comparteix amb el secessionista la tria sistemàtica de la variant lèxica valenciana quan aquesta contrasta amb la resta de la llengua, però accepta el sistema ortogràfic de base fabriana.
- Convergent: tendeix a usar formes comunes amb altres varietats de la llengua catalana, però al costat d'aquelles pròpies del valencià avalades també per una tradició escrita secular
- Uniformista: presenta formes com els subjuntius en -i ¸ possessius femenins amb -v-, o mots com noi, tarda, etc., que no empren la resta de models.

Encara que els quatre models tenen, o han tingut en el passat, algun tipus de suport polític, els estàndards principals al País Valencià es basen en els models de l'Institut d'Estudis Catalans (IEC), al que podríem anomenar *convergent* i el de l'Acadèmia Valenciana de la Llengua (AVL), que, encara que comparteix la mateixa base, presenta algunes diferències que l'acosten al *particularisme* (Sentí, 2019).

Com a exemples pràctics d'aquesta variació particular trobem els *Criteris lingüístics per als usos institucionals de les universitats valencianes,*[2] basats en el model de l'IEC i els *Criteris lingüístics de l'administració de la Generalitat,*[3] més propers als de l'AVL. Encara que tots dos comparteixen una gran part del seu contingut, també presenten algunes diferències:

[2] https://sl.ua.es/es/assessorament/documentos/criteris-linguistics.pdf

[3] https://ceice.gva.es/documents/161863132/163843714/Criteris+Ling%C3%BC%C3%ADstics_web.pdf

- Els criteris de les universitats accepten dues formes de l'imperfet de subjuntiu (amb -r- i amb -s-), mentre que els de la Generalitat només recomanen les formes amb -r-.
- Els universitaris recomanen usar avui i vuit en lloc de hui i huit, mentre que als criteris de la Generalitat és a l'inrevés.
- Els criteris de la Generalitat recomanen utilitzar formes d'infinitiu acabades en -re com *tindre* i *vindre*, mentre que els universitaris recomanen les formes *tenir* i *venir*.

Pel que fa a aquest treball, l'objectiu principal ha sigut automatitzar la quantitat més gran d'operacions de postedició de la manera més segura possible, evitant crear errors addicionals. És per això que les substitucions, encara que es podrien classificar dins del marc d'un model convergent, han estat guiades per la facilitat de la seua implementació amb relació a la seua freqüència. Per això podem trobar canvis com *feina→faena,* que podria caure dins del particularisme, però en cap cas trobarem substitucions corresponents a l'estàndard secessionista. És aquest també el motiu que fa que no hagen estat recollides diferències emblemàtiques des d'un punt de vista lingüístic entre les varietats orientals i occidentals del català com *llombrígol–melic* o *espill–mirall*, que es presenten en freqüències molt baixes.

1.3 Antecedents

Encara que l'objectiu del present estudi és una mica específic, hi ha alguns treballs relacionats que poden ser d'interés.

Guzmán (2007, 2008) fa en dues publicacions una exploració de l'ús de les expressions regulars com a mètode per automatitzar la postedició de l'eixida d'un traductor automàtic basat en regles. Després d'elaborar expressions regulars per corregir errors lingüístics recurrents, l'autor conclou que és possible automatitzar en gran manera la postedició de la traducció automàtica, cosa que permet reduir la

postedició manual i el cost de manera significativa. També afig que aquest tipus de postedició funciona millor en traductors automàtics basats en regles, ja que aquests traductors produeixen errors lingüístics de manera més consistent, en contraposició als sistemes de traducció estadístics. Encara que l'objectiu del present treball és també posteditar, com les substitucions es basen en les diferències entre estàndards lingüístics i no en errors de traducció, el fet que siguen traductors automàtics estadístics o basats en regles és irrellevant. A més, com cal substituir patrons que es repeteixen constantment, l'ús d'expressions regulars pot funcionar igual que en el cas de l'estudi del qual acabem de parlar.

Dins de la mateixa línia, Porro, Gerlach, Bouillon i Seretan (2014) també exploren la postedició automàtica, en aquest cas a partir de normes per corregir errors freqüents de la traducció al francés. L'estudi també conclou que l'automatització de la postedició amb normes com en el cas del seu projecte és beneficiós per a la posterior postedició manual.

Pel que fa a l'adaptació de diferents variants d'una mateixa llengua, trobem *AltLang,*[4] un convertidor automàtic basat en regles que treballa amb les varietats d'anglés americà i britànic, espanyol europeu i americà, francés europeu i canadenc i portugués europeu i brasiler. La tecnologia en aquest cas es basa en la plataforma de codi obert *Apertium* (Forcada, et al., 2011).

En últim lloc, cal esmentar el projecte d'adaptador valencià de Softcatalà,[5] amb un objectiu similar al d'aquest treball i que actualment compta amb una "versió web d'adaptació de textos poc revisada". Aquest adaptador usa el programa "sed", un programa que edita fluxos de text ("stream editor") i, entre d'altres, té operacions per a substitucions relativament similars a les que es presenten en el nostre treball. No obstant això, aquest adaptador, a diferència del present estudi, no permet tractar textos amb format.

[4] https://www.altlang.net/inicio/
[5] https://www.softcatala.org/projectes/adaptador-valencia/

2. Metodologia

2.1 Corpus d'entrenament

El primer pas per poder aconseguir els objectius del present estudi era identificar les operacions de substitució més comunes que calia fer per posteditar un text en català oriental per a "valencianitzar-lo". Per tant, es va decidir recopilar un corpus d'entrenament extens per veure quins eren els elements més comuns a substituir.

2.1.1 Creació del corpus

Per obtenir un corpus d'un àmbit el més general possible es va optar per recollir textos periodístics, extrets de l'arxiu web del diari *El País*. Es van seleccionar manualment notícies de temàtica variada, de diferents edicions de l'any 2015, fins a arribar a les 100.000 paraules.

Per tal de gestionar el corpus, tant mentre es recollia com una vegada ja acabat es va fer ús d'una eina fonamental per al treball en aquest projecte: *Git Bash*,[6] una aplicació per sistemes de Microsoft Windows que emula un entorn on poder fer ús de comandes procedents d'un sistema operatiu GNU/Linux. Això ens permet utilitzar les senzilles eines descrites a *Unix for Poets* (Church, 1994) per gestionar text.

En primer lloc, per a contar el nombre de paraules i veure com avançava el corpus, es va fer ús de la següent línia de comandament:

```
wc -w Corpus1ES.txt
```

[6] https://git-scm.com/downloads

On `wc` (*word count* en anglés) s'empra per a contar paraules, línies o caràcters, entre altres. `-w` especifica que són les paraules les que s'han de contar i `Corpus1ES.txt` indica el nom del fitxer on s'està treballant, en aquest cas, on s'estan recollint els textos de les notícies i la seua adreça web.

Pel que fa a les adreces web que acabem d'esmentar, s'apegaven al document en línies separades i precedides del símbol #, cosa que ens ajuda, una vegada completat el corpus, a separar-les de la resta del text. Per fer això, es van usar les següents instruccions:

```
grep '^#' <Corpus1ES.txt >Corpus1ESBi.txt
```

Primer, aquesta ordre usa la instrucció `grep` (acrònim de l'expressió anglesa *global regular expression print*) i constitueix la primera presa de contacte amb les expressions regulars i és que `'^#'` constitueix un exemple d'expressió regular. Aquesta instrucció ens permet cercar text,`'^#'` indica que han de ser línies de text que comencen per # (^ en aquest cas simbolitza el començament d'una línia), `<Corpus1ES.txt` és l'arxiu on ho ha de buscar i `>Corpus1ESBi.txt` és un nou fitxer de text on imprimirà el text recollit. És a dir, aquesta línia de comandament recull totes les adreces web de les notícies del corpus en un nou fitxer.

```
grep -v '^#' <Corpus1ES.txt >Corpus1ESOK.txt
```

Després, aquesta altra ordre només es diferencia de l'anterior per l'addició de `-v` que vol dir que imprimirà solament les línies de text que no coincideixen amb el patró (mateix d'abans, línies que comencen per #) i pel nom del fitxer d'eixida: `>Corpus1ESOK.txt`. Amb això aconseguim un fitxer nou del qual queden eliminades les adreces web de les notícies.

Una vegada ha quedat solament el text que ens interessa per continuar el treball el recompte definitiu de paraules indica que hi ha 103.139, així que supera per poc l'objectiu inicial de 100.000. Ara es pot procedir al següent pas: traduir-lo al català mitjançant Google Translate i analitzar els canvis que s'haurien de fer per posteditar el text per a valencianitzar-lo.

Com que el traductor de Google només ens deixa traduir com a molt 5000 caràcters en text pla de colp, es va procedir a passar el corpus a format .docx i així si es va poder traduir íntegrament mitjançant aquesta plataforma.

Finalment, cal destacar que les substitucions que es faran per valencianitzar el text seran de mots solts, com que el temps que caldria per a identificar i ordenar per la seua freqüència d'aparició expressions i conjunts de mots és massa elevat, queda fora de l'abast del present estudi.

2.1.2 Anàlisi del corpus

Amb el corpus d'entrenament ja traduït, era necessari trobar les operacions de postedició més freqüents. Amb aquesta finalitat, es va procedir a separar el corpus traduït automàticament en mots i ordenar-les segons la seua freqüència. Novament, això es va fer a través d'ordres en el mateix entorn que abans. Es va emprar la següent línia de comandament:

```
tr ' ' '\n' <Corpus1CAOK.txt | sort -f | uniq-c | sort -nr
>Corpus1CAPa.txt
```

Aquesta ordre funciona de la següent manera: `tr ' ' '\n'` separa les paraules del text, en substituir els espais en blanc (`' '`) per salts de línia (`'\n'`). Els símbols `<` i `>` indiquen, respectivament, d'on s'està extraent el text en el qual s'està treballant i el nom de l'arxiu on es guardarà el resultat. La barra vertical (`|`) o canonada serveix per a enviar l'eixida estàndard de l'ordre de l'esquerra a l'entrada

estàndard de l'ordre de la dreta, `sort -f` ordena els mots alfabèticament sense tenir en compte les majúscules, `uniq` agrupa qualsevol seqüència contigua de línies idèntiques en una única línia i `-c` indica quantes se n'han agrupat. A continuació, `sort -nr` els torna a ordenar, en aquest cas per ordre numèric (`-n`) invers (`r`), és a dir, per freqüència. Cal afegir que aquesta ordre tan senzilla no separa la puntuació, així que llista separadament, per exemple "feina" i "feina.", però no suposarà un problema en aquest treball, ja que la llista serà repassada manualment per identificar els mots que cal canviar. Ací es poden veure, per exemple, les cinc primeres entrades de la llista:

```
5047 de
3494 la
3131 que
2815 a
2362 i
```

Com es pot comprovar ací, els mots van acompanyades del nombre de vegades que apareixen al corpus. El següent pas és llavors analitzar els mots més freqüents que s'haurien de posteditar en cas de voler adaptar el text al valencià.

Per tal de seleccionar mots que havien de ser potencialment substituïts, es va fer una inspecció manual de la llista de mots, anotant qualsevol terme que poguera entrar dins d'aquesta descripció, a partir d'aparéixer en el corpus més d'una vegada i prestant especial atenció a aquelles que estaven més amunt en la llista.

Els criteris d'elecció van ser els següents:

- Formes verbals que solen canviar als estàndards valencians:
    - 1a Persona del singular del present d'indicatiu acabada en -o (*sento, parlo*).
    - Formes del present de subjuntiu (*acabi, arribis, parlin*).
    - Formes de l'imperfet de subjuntiu (*veiés, traguessis*).

- Altre vocabulari amb equivalents valencians normatius i usats (*mirall, feina*).
- Les formes femenines dels adjectius possessius (*meva, seva*)
- L'adverbi de lloc *aquí*.

A l'hora d'elaborar aquests criteris, com a norma general, s'ha tractat de no seleccionar casos acceptats pels estàndards valencians i usats amplament encara que tinguen alternatives pròpies valencianes, com canviar la terminació dels verbs incoatius (*compleix→complix*). Principalment per no carregar massa el treball que fem i perquè puga servir de base a la qual poder afegir altres substitucions depenent del criteri de qui ho vulga utilitzar en un futur.

Després de la selecció manual, tenim una llista de mots que cauen dins d'alguna de les categories que s'acaben d'esmentar. En total són 96 termes i a continuació es mostren els 5 més freqüents:

Paraula	Freqüència	Descripció[7]	Categoria gramatical	Substitució
seva	297	Pos F 3P Sg	Adjectiu	seua
seves	65	Pos F 3P Pl	Adjectiu	seues
sigui	47	Pr Sub 1o3P Sg	Verb	siga
feina	42	Lèxic	Nom	faena
cop	37	Lèxic	Nom	colp

Taula 1: 5 termes a substituir més freqüents al corpus d'entrenament

Dels 96 termes recollits, una gran majoria són verbs (80), després noms (11), adjectius (4) i finalment, adverbis (1).

[7] "Pos" = possessiu, "F" = femení, "P" = persona, "Sg" = singular, "Pl" = plural, "Pr" = present, "Sub" = subjuntiu, "Imp"=imperfet.

	Paraules a substituir
Adjectius	4
Noms	11
Adverbis	1
Verbs 1a persona singular present indicatiu	8
Verbs 1a/3a persona singular present subjuntiu	38
Verbs 2a persona singular present subjuntiu	1
Verbs 3a persona plural present subjuntiu	20
Verbs 1a/3a persona singular imperfet subjuntiu	9
Verbs 3a persona plural imperfet subjuntiu	4

Taula 2: classificació gramatical dels termes a substituir

- Dels verbs:
    - 8 són casos de la primera persona del singular del present d'indicatiu.
    - 72 són casos del subjuntiu:
        - 59 del present de subjuntiu: 38 de la 1a o 3a persona del singular; 1 de la segona persona del singular i 20 de la 3a persona del plural
        - 13 de l'imperfet de subjuntiu: 9 de la 1a o 3a persona del singular i 4 de la 3a persona del plural.
- Els noms constitueixen lèxic amb variants valencianes.
- Els quatre adjectius són les formes femenines dels possessius.
- L'adverbi de lloc *aquí*.[8]

Basant-nos en aquestes dades podem destacar l'aclaparadora majoria de verbs respecte a altres tipus de paraula. En canvi, si ens fixem en la freqüència, en les

[8] Encara que poc usat, és correcte en tots els estàndards per a referir-se al lloc de l'interlocutor. No obstant això, en els textos tots els *aquí* es refereixen al lloc del parlant i cal canviar-los per *ací*.

primeres cinc entrades s'inverteix la situació i són els noms i adjectius els que prevalen. Per tant, encara que hi ha recollits més casos diferents de verbs, la freqüència de les altres categories gramaticals compensa aquesta falta de diversitat per fer que els casos totals siguen similars.

2.2 Elaboració de les operacions de substitució

Era l'opció original d'aquest estudi utilitzar, de la mateixa manera que s'havia fet per a administrar els corpus, eines procedents del sistema operatiu GNU/Linux per elaborar les operacions de substitució mitjançant expressions regulars. No obstant això, finalment es va optar per usar les macroinstruccions del processador de textos LibreOffice.

Si el comparem amb l'altre sistema, les macroinstruccions ofereixen l'avantatge de treballar directament sobre textos amb un format determinat, com és habitual a l'àmbit dels processadors de text. A més, LibreOffice és programari lliure, potent, gratuït i de codi obert i, per tant, accessible per a la gran majoria d'ordinadors. D'altra banda, tenim l'inconvenient que el llenguatge que gasta és diferent i, en conseqüència, és necessari l'aprenentatge (encara que superficial) d'aquest.

Centrant-nos ara en les operacions de substitució, de la mateixa manera que es van classificar els termes de la llista de paraules a substituir, també s'han agrupat d'una manera similar les fórmules de substitució. Primer, però, hem d'observar la primera part de la macroinstrucció, que ens possibilita la mateixa substitució:

```
Sub Replacer(S as string, T as string)
RD = ThisComponent.createReplaceDescriptor()
RD.SearchRegularexpression = True
RD.SearchString = S
RD.ReplaceString = T
RD.SearchCaseSensitive = false
ThisComponent.ReplaceAll(RD)
End Sub
```

Aquest tros de codi defineix una "subrutina" (de fet una funció) "Replacer" que pren dos arguments, una cadena de caràcters d'entrada S que representa el text a substituir i una cadena de caràcters d'entrada `T` que representa el text que el substituirà. Definim un objecte `RD` que és un "Replace Descriptor", que es crea a partir de l'objecte "`ThisComponent`" que representa el document. Definim els valors d'alguns dels camps d'aquest objecte (el booleà `SearchRegularexpression` es defineix com a veritable, les cadenes `SearchString` i `ReplaceString` reben les cadenes que ha rebut el `Replacer`, i el booleà `SearchCaseSensitive` es fixa a fals perquè considere iguals les majúscules i les minúscules). Finalment, abans de retornar (`End Sub`), la funció invoca el mètode `ReplaceAll` de `ThisComponent` perquè execute la substitució que s'ha definit.

Pel que fa a les mateixes operacions de substitució, les trobem a una altra macro, agrupades en quatre grups:

- Substitucions de verbs en present de subjuntiu.
- Substitucions de verbs en imperfet de subjuntiu.
- Substitucions de verbs en primera persona del singular del present de subjuntiu.
- Altres substitucions de lèxic: ací trobem noms, adjectius i adverbis agrupats.

Una altra cosa que caldria comentar és el fet que es faran substitucions que en algun cas poden introduir errors, tanmateix, això només passarà en els casos on aquest efecte negatiu siga mínim

Abans d'endinsar-nos en les mateixes operacions, però, trobem peces de codi que, ja siga per motiu de la seua longitud o de la seua repetició en múltiples operacions s'han convertit en components que usaran les operacions. Trobem, principalment, variacions de `pre` i `post` que són *lookbehind* (`?<=`), que defineix el context previ a la cadena que cal substituir, però no se substitueix, i *lookahead* (`?=`),

que defineix el context posterior a la cadena que cal substituir, però no se substitueix. Aquestes parts són necessàries perquè funcionen bé les substitucions (sense elles hi havia problemes si les paraules a substituir se situaven a principi o final de línia o si hi havia dues juntes) i, a més, assegurar-nos que les substitucions es fan només on, per exemple, trobem mots complets.

Junt amb aquestes variacions, també hi ha llistes de verbs pels casos en els quals és necessari substituir els verbs cas per cas i no per la seua terminació. Les mostrem a continuació:

```
verbsI="(acab|apagu|apost|aprov|arrib|colpeg|comenc|consider|
culmin|dediqu|defens|deix|despagu|doblegu|don|estabilitz|expr
ess|figur|form|govern|marqu|necessit|pagu|pass|pens|permet|po
rt|pos|present|qued|ratifiqu|reb|renov|sembl|torn|torn|tract)
"
```

`verbsi` mostra les arrels de verbs que poden substituir "-i" per "-e" per canviar el present de subjuntiu.

```
verbsO="(don|imagin|intent|lament|llev|qued|recoman|record)"
```

`verbsO` recull arrels de verbs que poden substituir "-o" per "-e" per canviar la primera persona del singular del present d'indicatiu.

```
verbsAra="(acab|apost|aprov|arrib|consider|culmin|defens|deix
|don|estabilitz|express|figur|form|govern|necessit|pass|pens|
port|pos|present|qued|renov|sembl|torn|torn|tract)"
```

`verbsAra` és una llista d'arrels de verbs que substitueixen "-i" per "-ara" per canviar la 1a i 3a persona del singular de l'imperfet de subjuntiu.

```
verbsEra="(conegu|digu|hagu|permet|tingu)"
```

Finalment, `verbsEra` mostra les arrels de verbs que substitueixen "-i" per "-era" per canviar la 1a i 3a persona del singular de l'imperfet de subjuntiu.

2.2.1 Present de subjuntiu

Començant pel primer grup, la primera i una de les operacions més importants és la següent:

```
Replacer(preParcial++verbsI+")i(?=(s|n)|"+postParcial+")","e")
```

Aquesta operació és la que ens serveix per a canviar la 1a, 2a i 3a persona del singular i la 3a persona del plural del present de subjuntiu dels verbs que trobem en una de les llistes de verbs (`verbsI`) que acabem d'esmentar. `Replacer`, com hem vist, és el component que permet fer les substitucions al text, `preParcial` i `verbsI` contenen el *lookbehind* i la llista de verbs, la `i` és realment la part del mot que es va a substituir, per la `e` que se situa a la fi de l'expressió. Finalment, la part `(?=(s|n)|"+postParcial+")` indica que el mot pot acabar en "s" o "n", a més d'haver-hi el *lookahead*. Dit d'altra manera, el que fa aquesta línia de comandament és buscar una seqüència de caràcters situada entre l'inici d'una línia de text o un caràcter no alfanumèric (espais, punts...) que corresponga amb qualsevol de les formes verbals de la llista (per exemple, si el verb *acabar* té la forma "acabi", en la llista tenim "acab") seguida de la lletra "i" i el final d'una línia, un caràcter no alfanumèric o les lletres "s" o "n". Una vegada la troba, substitueix la "i" per una "e". Per tant, si troba "acabi", "acabis" o "acabin", ho substituirà per "acabe", "acabes" o "acaben".

A continuació, trobem:

```
Replacer("(?<=(^|[^g][aeiou]))eixi"+postSencer,"ïsca")
Replacer("(?<=(^||[^aeiou]))eixi"+postSencer, "isca")
Replacer("(?<=(^|[^g][aeiou]))eixin"+postSencer, "ïsquen")
Replacer("(?<=(^||[^aeiou]))eixin"+postSencer, "isquen")
```

Ací, aquestes quatre línies funcionen de manera similar a l'anterior. En aquests casos la diferència està en el fet que en lloc d'anar precedides d'un caràcter no alfanumèric seguit per un verb o una seqüència específica, qualsevol mot acabat en "-eixi" o "-eixin" canviarà aquestes terminacions per "isca/ïsca" o "isquen/ïsquen", depenent de si les terminacions van precedides de vocal `[aeiou]` o no `[^aeiou]`. A aquestes operacions també podem observar el *lookbehind*, perquè no s'ha utilitzat cap variació de `pre`. Funciona així: per tal de fer la substitució, la seqüència de caràcters que deu substituir-se ha d'estar precedida pel que queda definit entre els parèntesis que van després de `?<=`. En aquest cas és: `^`, que ací indica l'inici d'una línia (a diferència de quan va entre claudàtors, llavors s'empra com a negació), `[^aeiou]` i `[aeiou]`, que ja hem comentat i, finalment, `[^g]`. S'ha de destacar aquesta part (`[^g]`), que evita que utilitzar les terminacions amb dièresi per substituir mots que porten *g* abans de vocal i així evitar possibles errades com la substitució de "segueixi" per "seguïsca".

```
Replacer(preParcial+"(fa))ci"+postSencer,"ça")
Replacer(preParcial+"(fac))i"+postNs,"e")
```

Continuem amb aquestes operacions, que ens serveixen per a substituir les formes del present de subjuntiu del verb *fer*, la 1a/3a persona del singular del qual és una forma irregular ("faci"→"faça") i és la que substitueix la primera línia. La segona substitueix la resta de formes del present de subjuntiu ("facis"→"faces", "facin"→"facen").

```
Replacer(preVerbsGi+"gi"+postSencer,"ja")
Replacer(preVerbsGi+"gi"+postNs,"ge")
```

Després tenim dues operacions que s'encarreguen de substituir les formes dels verbs *anar*, *haver* i *veure*, que també són casos excepcionals quant al fet que per substituir la primera i tercera persona del singular cal canviar "-gi" per "-ja". Com abans, la segona operació s'encarrega de la resta de formes del present de subjuntiu d'aquests tres verbs.

```
Replacer("gui"+postSencer,"ga")
Replacer("gui"+postNs,"gue")
```

Aquestes operacions són una mica més arriscades, ja que substitueixen qualsevol mot acabat en "-gui", "-guis" o "-guin" per l'equivalent valencià corresponent. Després d'utilitzar eines web[9] per comprovar si hi podia haver problemes amb altres mots, els resultats indiquen en aquest cas que els possibles mots que es podrien substituir de manera errònia són pocs i poc freqüents (*iogui*). A més, els casos més comuns en els quals "-gui" s'haja de substituir per "-gue" també a la 1a i 3a persona del singular ("pagui" per "pague") estan coberts per la primera operació de substitució d'aquest grup (les substitucions es fan seqüencialment quan s'aplica la macroinstrucció). Per tant, aquestes instruccions disminuirien potencialment el temps de postedició en resoldre més problemes dels que creen.

```
Replacer("ï"+postsencer,"e")
Replacer("ï"+postNs,"e")
```

[9] https://www.paraulesamb.com/

Per últim, aquestes operacions, també un poc arriscades, substitueixen tot terme acabat en “-ï”, “-ïs” o “-ïn” per “-e”, “-es” o “-en”. D’una manera similar a les anteriors, els possibles problemes haurien de ser bastant inferiors a les substitucions correctes que realitzen. Encara que no s’ha trobat cap exemple concret on aquestes operacions produïsquen errades, sempre pot haver-hi algun nom propi que acabe en alguna de les terminacions substituïdes.

2.2.2 Imperfet de subjuntiu

Comencem per les següents operacions de substitució:

```
Replacer(preParcial+"(f[oe]))s"+postSencer,"ra")
Replacer(preParcial+"(f[eéoó]))ssi(?=(s|n|m|u)"+postParcial+"
)","re")
```

Ací tenim unes instruccions que substitueixen casos irregulars, concretament les formes dels verbs *fer* i *ser*. En el cas de la 1a i 3a persona del singular, substitueix la “-s” de “fos” i “fes” per “-ra”. Pel que fa a la resta, canvia “-ssi-” per “-re-“, així que, per exemple, “féssim” se substituiria per “férem”. Cal destacar que la segona operació sobregenera, és a dir, admet "fossiu" o "fóssis", però aquests mots són improbables i quedaran amb l'error en l'accent. A més, `preParcial` deixa el parèntesi del *lookbehind* obert i cal tancar-lo per tal que es complete, motiu pel qual trobem doble parentesi a `(f[oe]))` i `(f[eéoó]))`.

```
Replacer("(gu[eé])ssi"+postNsmu,"$1re")
```

Aquesta operació funciona de manera quasi idèntica a l’anterior, però s’aplica a totes les formes verbals que contenen en “-guessi-“, és a dir, 2a persona del singular i totes les formes del plural de l'imperfet de subjuntiu en verbs com *pagar* o *voler*. Com que aquesta operació no conté *lookbehind*, com que no és necessari pel fet que

es tracta del final d'un mot i no el mot sencer, cal col·locar `$1`, que representa el material dins dels primers parèntesis, abans de la substitució, per a evitar que consumisca aquesta part. Aquest problema es deu en part al fet que la implementació actual de *ReplaceAll* no permet combinar *lookahead* o *lookbehind* amb substitucions amb $ que conserven parts de la cadena que també van definides per expressions regulars.

```
Replacer(preParcial+"(don))essi"+postNs,"are")
Replacer(preParcial+"(don))éssi"+postMu,"àre")
```

Continuem amb aquestes instruccions, per al verb *donar*, de nou segueix una estructura similar a les anteriors, amb la diferència que la "e" que va davant de "ssi" s'ha de substituir per una "a", així que s'ha de fer en dues operacions perquè hi ha dues substitucions depenent de si s'accentua la "a" (1a i 2a persona del plural) o no (2a persona del singular i 1a del plural).

```
Replacer(preParcial++verbsAra+")és"+postSencer,"ara")
Replacer(preParcial++verbsEra+")és"+postSencer,"era")
```

Per finalitzar aquest grup, ací podem trobar dues operacions quasi idèntiques, que substitueixen les formes de la 1a i 3a persona del singular, substituint "-és" per "-ara" o "-era" en funció del verb, que en aquests casos tornen a estar recollits a dues llistes (`verbsAra` i `verbsEra`). Ací podem veure que com hi ha dos components junts `preParcial` i `verbsAra`, s'usa `+` per a enllaçar-los a la resta de l'operació i entre ells(`++`).

2.2.3 Primera persona del singular del present d'indicatiu

Aquest és el grup més curt, així que només trobem dues operacions de substitució:

```
Replacer(preParcial+"(s))ento"+postSencer,"ent")
```

La primera està feta per al cas del verb *sentir*, que perd la "-o" final de la varietat oriental. Ací se substitueix tota la forma llevat de la primera lletra perquè no produïsca errades amb les majúscules.

```
Replacer(preParcial++verbsO+")o"+postSencer,"e")
```

Com que durant l'elaboració del treball no es va descobrir cap patró que poder substituir amb més o menys èxit per a aquest cas, la solució que s'ha emprat ha sigut substituir la "-o" final per una "-e" als casos recollits en l'última llista de verbs que hi trobem a la macroinstrucció.

2.2.4 Altres substitucions

La primera operació d'aquest grup és la següent:

```
Replacer(preParcial+"[mts])ev(?=(a|es)"+postParcial+")","eu")
```

Amb la instrucció que tenim ací, qualsevol de les formes del femení dels adjectius possessius canviarà a la versió que se sol utilitzar al País Valencià. Funciona identificant mots que comencen per "m-", "t-" o "s-", seguits de "-ev-" i acaben en "-a" o "-es" i substitueix la part de "-ev-" per "-eu-". Així, per exemple, "meva", "seva" i "teves" passen a ser "meua", "seua" i "teues". Només amb aquesta línia

substituïm una gran quantitat de casos de la llista de mots, tenint en compte que només *seva* representa 297 casos, molts més que la segona entrada (*seves*) a la qual també cobreix aquesta operació.

```
Replacer(preSencer+"nen(?=(s|a|es)|"+postParcial+")","xiquet")
Replacer(preSencer+"noi(?=(s|a)|"+postParcial+")","xic")
Replacer(preSencer+"noies"+postSencer,"xiques")
```

Continuem ací amb tres operacions de substitució que canvien els termes típics orientals *nen* i *noi* per les formes més usades pel valencià *xiquet* i *xic*, amb les seues variacions en gènere i nombre. A la primera operació es cobreixen totes les formes de *nen*, gràcies al fet que l'equivalent d'aquest (*xiquet*) no canvia en cap de les formes ("xiquet**a**/**s**/**es**"). En canvi, per la forma femenina plural de *xic* (*xiques*) sí que calia fer una altra operació de substitució, pel fet que d'haver emprat la mateixa solució que en la primera operació hauria substituït "noies" per "xices".

```
Replacer(preSencer+"tarda"+postSencer,"vesprada")
```

Ací podem veure una altra operació que podríem descriure com arriscada, per substituir la paraula *tarda* per *vesprada*. Com sabem, a més d'indicar una part del dia, tarda és una forma verbal del verb *tardar*. No obstant això, com en el cas d'altres operacions, és prou més comú l'ús del terme *tarda* com equivalent de *vesprada*, que com a forma del verb *tardar*, especialment si tenim en compte que se sol utilitzar més el verb *trigar* amb aquest significat al català oriental.

```
Replacer(preSencer+"cop(?=(s)|"+postParcial+")","colp")
Replacer(preSencer+"sortid(?=(a|es)"+postParcial+")","eixid")
Replacer(preSencer+"patat(?=(a|es)"+postParcial+")","creïll")
```

```
Replacer(preSencer+"fein(?=(a|es)"+postParcial+")","faen")
Replacer(preSencer+"cruïlla"+postSencer,"encreuament")
Replacer(preSencer+"cruïlles"+postSencer,"encreuaments")
Replacer(preParcial+"(a))quí"+postSencer,"cí")
```

Finalment, s'agrupen ací la resta d'operacions de substitució de l'últim grup pel fet que funcionen de la mateixa manera, substitueixen mot per mot, amb variació del plural. Cada terme ocupa una línia a excepció de *cruïlla*, per la diferència en la forma del plural. A més, tots són noms menys per *aquí*, l'únic adverbi de totes les operacions.

I ací conclouen les operacions. Després de comprovar que funcionen correctament només queda veure els resultats.

3. Resultats

Per veure si el que hem fet fins ara compleix amb el seu objectiu, s'ha de provar el "prevalencianitzador" en quantitats de text suficients.

3.1 Prova en corpus

Amb aquest propòsit s'ha recopilat un segon corpus, aquesta vegada d'unes 10.000 paraules, del mateix origen de l'anterior (excepte pel fet que les notícies que el conformen són d'enguany) i recollit de la mateixa manera. Després, també com l'anterior, s'ha traduït al català mitjançant Google Translate.

Ja amb aquest corpus llest, es va apegar en un document en blanc al LibreOffice i es va executar la macroinstrucció. El temps d'execució va ser d'uns pocs segons, per la qual cosa podem afirmar que el rendiment no suposa un problema.

El següent pas, ara que tenim el text "prevalencianitzat" és comparar-lo amb la traducció original al català. El que trobem després d'una comparació manual en

l'entorn de l'eina en línia *diffchecker*,[10] és que s'han produït 128 canvis, distribuïts com podem veure en aquesta taula:

Terme	**Freqüència**	**Descripció**[11]	**Categoria gramatical**	**Terme substituït**
seua	60	Pos f 3a sg	Adjectiu	seva
seues	15	Pos F 3P Pl	Adjectiu	seves
faena	11	Lèxic	Nom	feina
ací	5	Lloc	Adverbi	aquí
meua	5	Pos F 1P Sg	Adjectiu	meva
meues	4	Pos F1P Pl	Adjectiu	meves
colp	2	Lèxic	Nom	cop
eixida	2	Lèxic	Nom	sortida
haja	2	Pr Sub 1o3P Sg	Verb	hagi
siga	2	Pr Sub 1o3P Sg	Verb	sigui
teua	2	Pos F 2P Sg	Adjectiu	teva
vesprada	2	Lèxic	Nom	tarda
acaben	1	Pr Sub 3P Pl	Verb	acabin
colps	1	Lèxic	Nom	cops
coneguera	1	Imp Sub 1o3P Sg	Verb	conegués
creïlles	1	Lèxic	Nom	patates
entenga	1	Pr Sub 1o3P Sg	Verb	entengui

[10] https://www.diffchecker.com/diff
[11] Vegeu la nota al peu 7

estiga	1	Pr Sub 1o3P Sg	Verb	estigui
faça	1	Pr Sub 1o3P Sg	Verb	faci
faenes	1	Lèxic	Nom	feines
fera	1	Pr Sub 1o3P Sg	Verb	faci
feren	1	Imp Sub 3P Pl	Verb	fessin
pogueren	1	Imp Sub 3P Pl	Verb	fes
puguen	1	Pr Sub 3P Pl	Verb	puguin
queden	1	Pr Sub 3P Pl	Verb	quedin
vinguen	1	Pr Sub 3P Pl	Verb	vinguin
xiquet	1	Lèxic	Nom	nen
xiquets	1	Lèxic	Nom	nens

Taula 3: Canvis al corpus de prova ordenats per freqüència

La conclusió que podem extraure d'aquestes dades és que, de la mateixa manera que al corpus d'entrenament, els mots més freqüents han estat les formes femenines dels adjectius possessius, especialment *seua* amb 60 substitucions, quasi un 50% del total. En general, la distribució dels termes substituïts és quasi idèntica a la de la llista de paraules a substituir.

Després d'analitzar els canvis un per un, cap de les 128 substitucions que s'han fet en aquest text de 10399 mots és errònia o causa problemes addicionals. Els únics casos en els quals hi podria haver alguna dificultat són ambdues substitucions de "cop" per "colp" i és que formen part de l'expressió "cada colp més", poc freqüent al País Valencià, on se sol emprar més "vegada" en lloc de "cop" en expressions similars. En principi, seria possible incloure aquesta expressió de més d'un mot amb els seus blancs en una de les operacions de substitució.

Pel que hem pogut veure, les substitucions han funcionat bé, però encara ens queda veure què ha quedat fora. No obstant això, el corpus és massa gran per comprovar-ho manualment. Així que, per veure què queda fora, s'ha decidit analitzar dos textos més, també periodístics, un d'unes 500 paraules i l'altre, més llarg, d'unes 2200. També s'han extret del diari *El País*.[12]

3.2 Altres textos

Aquests textos, com que són més curts, s'ha optat per posteditar-los manualment i al mateix temps passar-los la macroinstrucció per comprovar el percentatge de canvis que cobreix aquesta. Començant pel primer, el més curt, observem que totes les substitucions, tant de la postedició manual com de la macroinstrucció, estan concentrades a l'últim paràgraf. Si ens fixem en els dos primers paràgrafs, no seria necessari cap canvi. D'altra banda, en l'últim paràgraf trobem les substitucions:

> Ningú no creu que **siga** fàcil, perquè es tracta de repartir més justament els costos d'una situació sobrevinguda per la guerra a Ucraïna, per la dependència europea del gas i el petroli russos i pels canvis que pateix l'economia global. Aquesta situació ha fet que els augments de costos s'**hagen** traslladat als preus de les importacions. Però hi ha evidències més que suficients que la inflació està, sobretot, afectant els més febles, com ha assenyalat recentment el Banc d'Espanya. Activar mesures contundents d'estalvi a la factura energètica, com es va proposar en el passat recent, i centrar tota l'actuació de l'Estat en els més afectats per la pujada de preus haurien de ser els pilars d'una política econòmica

[12] https://elpais.com/opinion/2022-06-02/repensar-la-salida-a-la-crisis.html; https://elpais.com/espana/comunidad-valenciana/2022-06-05/el-rio-de-la-vida-en-valencia-un-extraordinario-jardin-interclasista-de-10-kilometros-y-35-anos.html

> que no s'acontentés amb estendre les mesures pal·liatives aprovades al març i esperar que amaini el temporal. Tot indica que, quan **arribe** la calma, la fallida social, el descontentament i la desafecció **puguen** ser irremeiables, si no s'actua abans.

Com podem observar, les quatre substitucions de la macroinstrucció (en negreta) són de formes del present de subjuntiu i són apropiades, mentre que les formes "acontentés" i "amaini" (subratllades) que sí han estat substituïdes a la postedició manual, han quedat fora. Això és perquè les operacions de substitució que haurien de processar-les depenen de fórmules amb llistes de verbs, i aquests verbs no hi apareixen. Encara que la mostra és massa petita per extraure conclusions generals, podem veure que, quatre dels sis canvis fets a la postedició manual han estat coberts, és a dir, un 66%.

Pel que fa al segon text, el que podem veure és ben curiós. La macroinstrucció ha realitzat 25 substitucions, enfront de les 29 de la postedició manual; totes les substitucions de la macroinstrucció han sigut de nou correctes, però, a diferència de l'anterior, només hi ha hagut una forma verbal que ha quedat fora, *baixo*, encara que no hi hauria de ser al text, pel fet que es tracta d'una errada de traducció automàtica ("bajo mandato del PP" passa a "baixo mandat del PP"). Hi ha un altre cas similar, i és que s'ha substituït *tarda* per *vesprada* per una altra errada de la traducció automàtica ("i més tarda" →"i més vesprada"). Pel que fa a la resta del text, les substitucions que deixa fora la macroinstrucció tenen a veure amb la traducció automàtica del text original de *cauce* per *llera* que, com que parla del Jardí del Túria, caldria substituir per *llit*. Podem entendre que com que el text fa poc ús de formes més conflictives com ara el subjuntiu o les primeres persones del singular del present d'indicatiu és per aquest motiu que només hi ha un mot que no s'ha substituït, encara que es repeteix en diferents ocasions. No obstant això sí trobem alguns exemples que entrarien dins d'aquestes categories que sí s'han substituït, com: "tinguessin"→"tingueren", "fos"→"fora", "pugui"→"puga" o "recordo"→"recorde".

En tot cas podem afirmar que, de nou, la macroinstrucció abasta la majoria de paraules a substituir, i és que la macroinstrucció ha cobert 23 (restem els dels errors de la traducció automàtica) dels 29 canvis de la postedició manual, un 79%.

4. Conclusions

L'objectiu del present estudi era el d'explorar la "prevalencianització" de l'eixida dels traductors automàtics al català mitjançant operacions de substitució basades en expressions regulars. Amb això com a finalitat, s'ha elaborat un corpus extens d'entrenament i s'ha extret d'aquest una llista de mots a canviar basats en la freqüència d'aparició. A partir d'aquesta llista s'han implementat les esmentades operacions de substitució basades en expressions regulars, dins del context de les macroinstruccions del processador de textos LibreOffice. Finalment, s'ha comprovat l'eficàcia d'aquestes instruccions mitjançant un altre corpus. La hipòtesi inicial era que amb aquestes eines podríem alleugerir la postedició de l'eixida dels traductors automàtics al català pel cas que es volguera adaptar a qualsevol dels estàndard del País Valencià. Els resultats ens ho confirmen, encara que amb algunes limitacions.

- Per una banda, com s'ha vist en l'anàlisi del text curt, hi ha formes que queden fora de l'abast de les operacions de la macroinstrucció i, a més, existeix la possibilitat que algunes de les operacions de substitució causen traduccions errònies, encara que siga en casos molt poc freqüents, com *tarda→vesprada*. A més a més, hi ha algunes operacions, com les que substitueixen termes sencers ("nen"-"xiquet") que no respecten l'ús de les majúscules del text original, encara que tècnicament seria possible resoldre aquest problema refinant algunes regles. Més enllà, també seria possible substituir, no només termes aïllats, sinó també expressions com "cada

cop més" o "casa seva", d'ús poc freqüent a terres valencianes, que actualment queden fora de l'abast del present estudi.

- Per l'altra, com hem vist als resultats, la macroinstrucció compleix la seua funció: ha fet un nombre considerable de substitucions amb un percentatge d'error mínim. Encara que no cobreix tots els casos possibles de substitució sí que ho fa d'una majoria. Pel que fa al rendiment, solament tarda uns segons a executar-se i, com que està implementada a un processador de textos, els textos poden mantenir el seu format. També és important destacar la flexibilitat del format i és que a partir de les eines i les fórmules de substitució ja desenvolupades es pot completar i ampliar la macroinstrucció perquè incloga els casos i formes que ara queden fora.

Caldria també parlar del motiu que hi haja poques errades es deu en part al fet que les substitucions suposen, en el cas del corpus, únicament menys de l'1,3% de les paraules. En part perquè, com hem dit abans, com que està basat en mots, hi ha expressions que són més naturals al català oriental que es podrien substituir per altres més usades al País Valencià. També per les paraules que queden fora, encara que pareixen ser menys que les que sí que són substituïdes. Així mateix, la diferència entre el català normatiu oriental i les normes valencianes no és massa gran, especialment si gastem l'estàndard de l'IEC.

En definitiva, com hem vist, l'elaboració d'operacions de substitució basades en expressions regulars ha resultat un èxit, encara que amb limitacions, per tal d'alleugerir la postedició de l'eixida dels traductors automàtics al català per adaptar-los als estàndards valencians.

Treball futur

Com s'ha parlat a les conclusions, per tal de millorar el que s'ha fet dins d'aquest treball es poden dur a terme els següents exemples:

- Un estudi d'expressions de més d'un mot i d'altres característiques lingüístiques típiques del català oriental i els seus equivalents valencians, per a poder-les implementar dins la macroinstrucció. Com en el cas de les substitucions de *cop*.
- Un refinament de la macroinstrucció per solucionar els possibles problemes amb majúscules.
- Una ampliació de les formes i casos que ja inclou la macroinstrucció per a poder reduir el nombre d'aquests que queden fora, amb cura de no alterar massa el rendiment. Per exemple, mitjançant les llistes de verbs.
- Una anàlisi més en profunditat del funcionament de la macro al seu estat actual.
- Queda també oberta la possibilitat de traslladar el projecte del llenguatge de programació BASIC, que usa actualment, a Python l'altre llenguatge de programació amb el qual es poden fer macroinstruccions al LibreOffice.

Referències bibliogràfiques

Church, K. W. (1994). Unix® for Poets. *Notes of a course from the European Summer School on Language and Speech Communication*.

Forcada, Mikel L.; Ginestí-Rosell, Mireia; Nordfalk, Jacob; O'Regan, Jim; Ortiz-Rojas, Sergio; Pérez-Ortiz, Juan Antonio; Sánchez-Martínez, Felipe; Ramírez-Sánchez, Gema; Tyers, Francis M. (2011). Apertium: a

free/open-source platform for rule-based machine translation. *Machine Translation*(25), 127-144.

Goyvaerts, J. (19 / Agost / 2021). Recollit de regular-expressions.info: https://www.regular-expressions.info/tutorial.html

Guzmán, R. (9 de 2007). Automating MT post-editing using regular expressions. *Multilingual*, 49-52.

Guzmán, R. (2008). Advanced automatic MT post-editing. *Multilingual Computing*, 52-27.

Mas, J. À. (2008). *El Morfema ideològic. Una anàlisi crítica dels models de llengua valencians.* Benicarló: Onada Edicions.

Mas, J. À. (2010). Les connotacions ideològiques dels models lingüístics valencians: situació actual i condicionants històrics. *Caplletra*, 47-70.

Porro, V., Gerlach, J., Bouillon, P., & Seretan, V. (2014). Rule-based Automatic Post-processing of SMT Output to Reduce Human Post-editing Effort. *Proceedings of the 36th International Conference on Translating and the Computer.* Londres.

Sentí, A. (2019). La (re)construcció de l'estàndard lèxic valencià: un estudi d'actituds. *Treballs de Sociolingüística Catalana*, 133-153.

Annex

Dins d'aquest apartat s'enllacen dues parts importants d'aquest treball perquè hi puga accedir qui vulga:

- La macroinstrucció que s'ha elaborat, juntament amb les instruccions per instal·lar-la està disponible al següent enllaç del repositori GitHub: https://github.com/reconco15/valencianitzador
- Els corpus i els textos amb els quals s'ha provat la macroinstrucció estan disponibles dins d'aquesta carpeta de lliure accés de Google Drive: https://is.gd/GbmvTv