

CONSTRUIR SISTEMES DE TRADUCCIÓ AUTOMÀTICA PER A LLENGÜES MENORS: REPTES I EFECTES

Mikel L. Forcada*

Resum

La creació de sistemes de traducció automàtica per a llengües desfavorides, que anomenaré llengües *menors*, presenta diversos reptes alhora que obri la porta a noves oportunitats. Després de definir conceptes preliminars com ara els de *llengua menor* i *traducció automàtica*, i d'explicar breument els tipus de traducció automàtica existents, els usos més comuns, el tipus de dades en què es basen, i els drets d'ús i les llicències del programari i de les dades de traducció automàtica, es discuteixen els reptes a què s'enfronta la construcció de sistemes de traducció automàtica i els possibles efectes sobre l'estatus de la llengua menor, usant com a exemples llengües menors d'Europa.


Paraules clau: traducció automàtica; llengües menors; recursos lingüístics; aragonés; bretó; sami; noruec *bokmål*; noruec *nynorsk*; occità; català; valencià.

BUILDING MACHINE TRANSLATION SYSTEMS FOR MINOR LANGUAGES: CHALLENGES AND EFFECTS

Abstract

Building machine translation systems for disadvantaged languages, which I will call minor languages, poses a number of challenges whilst also opening the door to new opportunities. After defining a few basic concepts, such as minor language and machine translation, the paper provides a brief overview of the types of machine translation available today, their most common uses, the type of data they are based on, and the usage rights and licences of machine translation software and data. Then, it describes the challenges involved in building machine translation systems, as well as the effects these systems can have on the status of minor languages. Finally, this is illustrated by drawing on examples from minor languages in Europe.

Keywords: *machine translation; minor languages; language resources; Aragonese; Breton; Saami; Norwegian Bokmål; Norwegian Nynorsk; Occitan; Catalan; Valencian.*

* Mikel L. Forcada, catedràtic de llenguatges i sistemes informàtics en la Universitat d'Alacant, president de l'Associació Europea de Traducció Automàtica (EAMT) des de 2015, fundador i president del Comitè de Gestió de la plataforma de traducció automàtica de codi obert Apertium, i cofundador i director d'investigació de l'empresa de tecnologia lingüística Prompsit Language Engineering. mlf@dlsi.ua.es  0000-0003-0843-6442

Article rebut el 14.01.2020. Avaluacions cegues: 26.02.2020 i 22.04.2020. Acceptació de la versió final: 22.05.2020.

Citació recomanada: Forcada, Mikel L. (2020). Construir sistemes de traducció automàtica per a llengües menors: reptes i efectes. *Revista de Llengua i Dret, Journal of Language and Law*, 73, 1-20. <https://doi.org/10.2436/rld.i73.2020.3404>

Sumari

- 1 Introducció
- 2 Definicions preliminars
 - 2.1 Llengües menors
 - 2.2 Traducció automàtica
 - 2.2.1 Definició
 - 2.2.2 Usos
 - 2.2.3 Tipus
 - 2.3 Dades
 - 2.3.1 Recursos lingüístics
 - 2.3.2 Corpus
 - 2.4 Drets d'ús i llicències
 - 2.4.1 Programari
 - 2.4.2 Dades
- 3 Reptes
 - 3.1 Actituds tecnofòbiques o luddites
 - 3.2 Estandardització
 - 3.3 Elicitació del coneixement i complexitat lingüística
 - 3.4 Organització de la creació dels recursos
 - 3.5 Gestió dels recursos: llicències i comuns
- 4 Efectes
 - 4.1 Normalitat i visibilitat
 - 4.2 Alfabetització
 - 4.3 Estandardització
 - 4.4 Increment de l'expertesa i dels recursos disponibles
- 5 Estudi de casos
 - 5.1 Bretó
 - 5.2 Occità
 - 5.3 Aragonés
 - 5.4 Traducció entre noruec *bokmål* i noruec *nynorsk*
 - 5.5 El traductor de sami del nord a noruec *bokmål*
 - 5.6 El traductor automàtic de la Generalitat Valenciana
- 6 Conclusions
- Referències bibliogràfiques

1 Introducció

Una bona part de la nostra vida és *en línia*. Això requereix eines per a processar textos eficientment; quan aquestes eines no són a l'abast d'una determinada llengua, els usuaris d'aquesta llengua han de canviar a una altra per a poder viure en línia i, com a resultat, la seua experiència de vida en línia és limitada. Una eina important per a processar els textos d'una llengua és la *traducció automàtica*. Aquest article descriu els reptes i els efectes de la construcció de sistemes de traducció automàtica per a llengües en situacions de desavantatge, que anomenarem *menors*.

L'article s'organitza així: l'apartat 2 defineix conceptes com ara *llengua menor* i *traducció automàtica*, explica breument els tipus de traducció automàtica existents, els usos més comuns, el tipus de dades en què es basen, i discuteix els drets d'ús i les llicències del programari i de les dades de traducció automàtica. L'apartat 3 descriu els reptes a què s'enfronta la construcció de sistemes de traducció automàtica i l'apartat 4 repassa els possibles efectes sobre l'estatus de la llengua menor. Abans dels comentaris finals, l'apartat 5 dona els exemples de sis llengües minoritzades d'Europa: el bretó, l'occità, l'aragonés, el noruec *nynorsk*, el sami septentrional i el d'una llengua no tan fortament minoritzada, el català.

2 Definicions preliminars

2.1 Llengües menors

Aquest article usa el concepte de *llengua menor* per a englobar diverses situacions de desavantatge de les llengües, com vaig fer en un altre article fa catorze anys (Forcada, 2006).¹ En la bibliografia i en Internet s'utilitzen moltes expressions diferents de manera més o menys intercanviable amb el de *llengua menor*. A més de la denominació *llengües menors*, he observat també les denominacions relacionades següents (amb el nombre de resultats de Google per al terme en català entre parèntesis): *llengües minoritàries* o llengües parlades per una minoria dins d'un estat o territori (29.900), *llengües menys usades*, és a dir, llengües amb un nombre d'usuaris menut comparat amb el d'altres (100), *llengües petites* (1.160), *llengües més petites* (892) *llengües amb pocs recursos* (229), etc. En aquest article no tindrè en compte les diferents connotacions de cada terme (per exemple, una llengua minoritària d'un país pot ser una llengua gran del món, com el gujarati al Regne Unit), sinó que simplement usaré el terme *llengua menor* (374) per referir-me a una llengua que mostre algunes, si no totes les següents característiques (vegeu Streiter et al., 2006):

- té un nombre reduït de parlants (o, ja que aquest article tracta de la traducció automàtica i, per tant, parlem de textos, un nombre reduït de parlants *alfabetitzats*);
- s'usa lluny de la normalitat (per exemple, s'usa més a casa o en situacions familiars que a l'escola, el comerç o l'administració, és discriminada socialment, abandonada políticament, mal finançada, prohibida o reprimida, etc.);
- no té un sistema d'escriptura únic, una ortografia estable o una varietat estàndard àmpliament acceptada;
- té una presència molt limitada a Internet;²
- té pocs lingüistes experts;
- fretura de recursos en suport informàtic llegibles per l'ordinador, com ara diccionaris, corpus, etc.;
- depèn de tecnologies que no són fàcilment accessibles als parlants.

En l'article, s'usen tèrmins com *llengües principals*, *llengües majors* o *llengües dominants* per a referir-nos a les llengües que no presenten cap d'aquestes limitacions.

¹ Parts d'aquest article són, de fet, una versió actualitzada d'aquell.

² O com ja deien Williams et al. (2001), és "no visible en la interactivitat natural mitjançada pels sistemes de l'era de la informació".

Considerem per exemple les llengües d'Europa, o més concretament de la Unió Europea (UE). La UE és —i vol ser, almenys en les declaracions oficials— multilingüe. Les identitats lingüístiques són al nucli de com els europeus es perceben a si mateixos i com entenen les seues relacions. La diversitat lingüística a Europa és un actiu molt valorat, però també és un repte i una causa de fragmentació. Una gran part dels europeus són (encara) funcionalment monolingües.³ Es tracta d'una possible font d'injustícia, ja que impedeix que la participació social plena en la societat europea s'estenga més allà de les seues barreres lingüístiques. De fet, es percep que els individus multilingües tenen un avantatge definit a Europa.⁴

Clarament, no tots els ciutadans d'Europa tenen els mateixos drets. Fins i tot dins d'un estat de la UE, els parlants d'una de les 24 llengües declarades oficials (per exemple, els parlants de neerlandés als Països Baixos) tenen més drets que els parlants d'una llengua no declarada oficial (per exemple, un bascoparlant a França). I quan es tracta de relacionar-se amb institucions de la UE, els parlants de llengües que no estan en la *llista del Tractat de Lisboa* troben impediments severos. Si tenim en compte que la comunicació és un component important en la construcció de la ciutadania, podem concloure que els ciutadans experimenten nivells de ciutadania diferents segons quina siga la seua llengua L_1 .

En l'actualitat, una bona part de les nostres vides té lloc *en línia*. En particular, com més va més interaccionem en línia amb institucions i empreses. Per tant, perquè una llengua siga útil, ha de ser útil en línia. En particular, això requereix que hi haja textos disponibles en aquesta llengua i que aquests textos es puguen processar en línia de manera eficient. Les tecnologies de la llengua són, per tant, una peça clau en la construcció de la ciutadania en l'actual societat digitalitzada.

Aquest article posa el focus sobre una de les tecnologies de la llengua: la traducció automàtica. Com que la traducció automàtica connecta les llengües entre elles, l'efecte de la disponibilitat de traducció automàtica per a una llengua menor M sempre es produirà en connexió amb altres llengües:

1. En connexió amb una llengua principal P que té relacions de traducció amb la llengua menor M , gràcies a l'existència d'un sistema de traducció automàtica entre M i P . Per exemple, M =bretó, P =francès.
2. En connexió amb una llengua principal P que té relacions de traducció amb una altra llengua N molt propera a la llengua M . Si dues llengües menors M i N són molt similars, serà més fàcil crear un sistema de traducció automàtica entre elles; si ja hi ha traducció automàtica d'un idioma gran P a una d'elles, N , l'altra llengua menor, M , pot beneficiar-se de l'existència de traducció automàtica —indirecta, a través de N — cap a aquesta. Per exemple, podem traduir indirectament de P =anglès a M =asturià a través de N =espanyol.

2.2 Traducció automàtica

2.2.1 Definició

La traducció automàtica tracta amb textos escrits, i, en particular, amb *textos informatitzats*, és a dir, amb documents de text emmagatzemats en un mitjà informàtic; és a dir, documents com els que es poden generar o editar amb processadors de textos. És *automàtica* perquè la realitzen *sistemes informàtics*, és a dir, ordinadors amb el programari adequat instal·lat, sense intervenció humana. Per tant, entenem per *traducció automàtica* la transformació, usant un sistema informàtic, d'un text informatitzat escrit en la *llengua origen* en un altre text informatitzat escrit en la *llengua meta* que podem anomenar *traducció en brut*.

La traducció automàtica té limitacions. En general, les traduccions en brut produïdes pels sistemes de traducció automàtica solen ser diferents de les produïdes pels professionals de la traducció i poden no ser adequades per a alguns propòsits comunicatius. Aquesta inadequació és causada per la dificultat a l'hora d'abordar automàticament diversos problemes, entre els quals podem comptar l'ambigüitat dels textos humans (que

³ En l'enquesta Eurobarometer Special (2012), el 46% dels europeus afirmen que poden mantenir una conversa només si és en la seua llengua materna.

⁴ Vegeu també l'estudi de Rivera et al. (2017) encarregat pel Panel pel Futur de la Ciència i la Tecnologia del Parlament Europeu.

contenen moltíssims mots amb més d'un sentit⁵ o frases amb més d'una possible estructura sintàctica),⁶ les divergències sintàctiques entre la llengua origen i la llengua meta,⁷ etc. Aquests problemes s'aborden amb mètodes que, en general, fan simplificacions bastant radicals del procés de traducció; simplificacions que, d'una banda, permeten la formulació de mecanismes de traducció prou senzills per a construir sistemes de traducció automàtica ràpids i compactes en un temps raonable, però que, de l'altra banda, fan que les solucions siguin lluny de ser òptimes.

En vista d'aquestes limitacions podem esperar que un bon sistema de traducció automàtica ens allibere de la part més mecànica (o "mcanitzable") de la tasca de traducció, però, per bo que siga, no podem esperar que compregui el text, resolga sempre les ambigüitats correctament i produïska textos en una variant genuïna de la llengua meta.

2.2.2 Usos

Hi ha dos grans grups d'aplicacions de la traducció automàtica.

El primer grup el formen les aplicacions per a l'*assimilació*, és a dir, l'ús de la traducció automàtica per a comprendre el sentit general dels documents (per exemple, textos publicats en Internet) escrits en altra llengua. Exemples:

- Una oficina de patents usa la traducció automàtica per exemple per a determinar si les patents en altres idiomes infringeixen les seues patents o si aquestes estan en perill d'infringir les patents en altres idiomes (Nurminen, 2019).
- Cada persona que participa en un *xat* pot escriure en la seua llengua i llegir les contribucions dels altres participants en la seua llengua gràcies a la integració d'un sistema de traducció automàtica en l'aplicació de *xat*.
- En les xarxes socials com ara Twitter o Facebook, un botó ens permet llegir publicacions en altres llengües traduïdes a la nostra llengua.

En aquest tipus d'aplicacions la traducció automàtica ha de ser molt ràpida, idealment instantània, i s'usa directament en brut; hi ha vegades que el resultat ni tan sols es llig completament (per exemple, quan el document traduït és molt llarg), i normalment no es conserva ni guarda després d'haver-lo llegit. Aquesta aplicació de la traducció automàtica no està pensada per a professionals de la traducció sinó per al públic en general.

En el segon grup, hi ha les aplicacions per a la *disseminació*, on la traducció automàtica l'usen professionals de la traducció. Es diuen *de disseminació* perquè comporten l'ús de la traducció automàtica com a pas intermedi en la producció d'un document en la llengua meta que serà publicat o disseminat; per tant, la traducció en brut normalment es conserva perquè l'ha de revisar i corregir, o com se sol dir, *posteditar*, una persona idealment especialitzada, professional de la traducció.⁸ Es pot traduir el document complet i després posteditar-lo en un processador de textos, però és més còmode integrar la traducció automàtica en un entorn de traducció assistida per ordinador on la traducció automàtica és una eina de suport, com també ho són les

5 Com ara el mot polisèmic espanyol *registrar*, que pot voler dir "inscriure en un registre" (en català *registrar*) o "examinar algú o alguna cosa per trobar alguna cosa que estiga oculta" (en català *escorcollar*, *inspeccionar*), o el mot homògraf espanyol *libertad*, que pot ser un nom (en català *llibertat*) o un verb en imperatiu (en català *allibereu*).

6 Com ara l'oració *Va comprar les taronges que va vendre a Empar* on el sintagma preposicional *a Empar* pot modificar el sintagma verbal de l'oració subordinada, *va vendre*, o el sintagma verbal de l'oració principal que el conté, *comprar les taronges que va vendre*. En el primer cas, Empar és compradora i en el segon, venedora.

7 L'oració basca *Donostiatik etorri den gizonari eman diot* (en català *Ho he donat a l'home que ha vingut de Donostia*), literalment és "Donostia-de vingut és-que home-el-a donat li-ho-he". La divergència sintàctica fa que la traducció comporte un reordenament radical dels elements de l'oració (en aquest cas, una inversió completa).

8 Com més va més s'usen models de proveïment participatiu (en anglés, *crowdsourcing*), amb afany de lucre o sense, per a generar contingut publicable a partir de la traducció automàtica, en els quals també intervenen no professionals de la traducció. En alguns casos, ni tan sols es tracta de postedició sinó d'edició monolingüe en la llengua meta.

memòries de traducció (que permeten recuperar traduccions fetes anteriorment per a oracions similars) o les bases de dades terminològiques (que contenen traduccions validades per als tèrmins d'especialitat) o les eines per a calcular costos i productivitat. En qualsevol cas, simplificant una miqueta, podem dir que la traducció automàtica seguida de postedició constitueix una alternativa a la traducció professional només si el seu cost conjunt és menor que el de la traducció professional tradicional, o quan es vol accelerar el procés de traducció mantenint el cost.⁹

2.2.3 Tipus

També s'hi poden distingir *dos grans grups de tecnologies de traducció*.

Des dels primers intents de fa uns 60 anys fins al decenni dels noranta, l'aproximació dominant a la traducció automàtica ha estat l'anomenada *traducció automàtica basada en regles* (TABR), que podem trobar en sistemes com ara Lucy,¹⁰ ProMT¹¹ o Apertium.¹² Típicament, la traducció automàtica basada en regles progressa a partir de la traducció mot a mot, idealment fins a tenir en compte tota l'oració. Per a desenvolupar un sistema de TABR:

- D'una banda, els experts en traducció compilen diccionaris en forma electrònica, escriuen regles que analitzen la llengua origen i que transformen estructures de la llengua origen en estructures equivalents de la llengua meta, etc. (noteu que el coneixement intuïtiu i no formalitzat dels traductors sobre la tasca s'ha de convertir en regles i s'ha de codificar de forma eficientment computable; això pot obligar a fer simplificacions radicals, que, si s'elegeixen bé, poden ser útils en la majoria dels casos).
- De l'altra banda, els experts informàtics escriuen programes (anomenats motors de traducció) que consulten els diccionaris i apliquen (en l'ordre previst) les regles al text original per a analitzar-lo i traduir-lo.

Més recentment, des de principis dels noranta assistim a un creixement de l'anomenada *traducció automàtica basada en corpus* (TABC): els programes de traducció automàtica "aprenen a traduir" a partir d'enormes corpus de textos bilingües on centenars de milers o milions de frases en una llengua s'han emparellat o *alineat* amb la seua traducció en l'altra llengua (és a dir, enormes *memòries de traducció*). En el cas de la traducció automàtica basada en corpus, el paper dels experts en traducció podria parèixer menys important si no es tinguera en compte l'esforç de traducció (idealment, però, no sempre professional) present en els corpus d'entrenament (vegeu l'apartat 2.3.2 sobre altres possibles fonts de corpus bilingües d'entrenament).

Hi ha dues estratègies principals de traducció automàtica (TA) basada en corpus: la TA *estadística* i la TA *neural*:

- La TA estadística, inventada a finals dels anys vuitanta, i que s'aplica comercialment aproximadament del 2003 ençà, aprèn i usa models probabilístics que s'estimen comptant determinats successos en el corpus bilingüe d'entrenament (per exemple, quantes voltes apareix un determinat mot al costat d'un altre mot determinat en l'oració meta, o quantes voltes apareix un mot determinat en l'oració origen quan un altre mot determinat apareix en l'oració meta).
- La nova TA neural s'explota comercialment des del 2016. Es basa en xarxes neurals artificials inspirades (vagament) en la manera com el cervell aprèn i generalitza; en aquest cas, aprenen i generalitzen a partir de l'observació dels corpus bilingües (Forcada, 2017; Casacuberta i Peris,

⁹ A voltes, per a estalviar postedició es pot fer una miqueta de preedició del text original que es traduirà automàticament, amb la qual cosa s'eviten problemes coneguts del sistema de traducció automàtica concret que s'estiga usant.

¹⁰ <http://lucysoftware.com/catala/traduccion-automatica/kwik-translator/>

¹¹ <https://www.online-translator.com/>

¹² <http://www.apertium.org>

2017). De fet, els principals sistemes públicament disponibles en línia de Google,¹³ Microsoft,¹⁴ etc. ja són neurals i hi ha a més, sistemes neurals nous com ara DeepL.¹⁵

Per descomptat, també hi ha sistemes *híbrids* que integren les dues estratègies (per exemple, usen regles morfològiques per a analitzar el text abans de traduir-lo usant un sistema entrenat també sobre un corpus de textos morfològicament analitzats).

Els sistemes de traducció automàtica basats en regles requereixen bastant faena lingüística i traductològica i més temps de construcció, ja que cal codificar explícitament les dades lingüístiques, és a dir, les regles i els diccionaris, de manera que les pugui usar el sistema; en canvi, els sistemes basats en corpus són més ràpids de construir, però només si prèviament es disposa d'un volum suficient de corpus de textos traduïts alineats oració a oració; per tant, poden ser difícils d'aplicar a la traducció d'una llengua minoritària amb poca disponibilitat de corpus digitalitzats. Això fa que, en aquest últim cas, la TABR pugui ser l'única estratègia amb possibilitats de reeixir. Per això aquest article hi farà inevitablement més èmfasi que sobre la TABC.¹⁶

És important tenir també en compte que els sistemes de TA estadístics i encara més els neurals poden produir textos enganyosament naturals —per la importància que donen a imitar els textos meta usats per a entrenar-los—, però que no són una traducció adequada del text original.

Independentment de quina siga la tecnologia concreta, podem dir que en un sistema de traducció automàtica es poden distingir tres components:

- un *motor*, és a dir, el programa que fa la traducció automàtica;
- unes *dades* (recursos lingüístics o corpus, com veurem en l'apartat següent); i
- les *eines* per gestionar i mantenir aquestes dades i convertir-les o transformar-les en el format requerit pel motor (en el cas dels sistemes basats en corpus, aquesta transformació inclou el procés d'aprenentatge dels models probabilístics o neurals corresponent).

2.3 Dades

Siga quin siga el tipus de traducció automàtica, no és possible si no es disposa de dades per al parell de llengües (origen i meta) en formats llegibles per l'ordinador. La naturalesa d'aquestes dades depèn del tipus de traducció automàtica, com s'explica de seguida. Distingirem dos tipus de dades en aquest article: d'una banda, els *recursos lingüístics*, i de l'altra, els *corpus*.

2.3.1 Recursos lingüístics

En el cas de la traducció automàtica basada en regles, hem de proveir el motor de *recursos lingüístics* (ens referim ací a recursos en un format llegible per l'ordinador, no necessàriament per persones) com ara diccionaris monolingües que descriuen la morfologia de les llengües origen i meta, regles per a desambiguar els mots homògrafs i polisèmics, regles per a transformar les estructures de la llengua origen a estructures equivalents de la llengua meta, diccionaris bilingües, etc. Aquests recursos han d'estar emmagatzemats en el format que esperen les *eines* i el mateix *motor* de traducció. Com s'ha dit més amunt, aquests recursos costen molt de construir, i requereixen la disponibilitat d'experts en lingüística i traducció familiaritzats amb els formats que use el sistema; els experts han de crear recursos completament nous o transformar recursos existents.

Els recursos lingüístics també es poden usar per a transformar, *anotar* o preparar automàticament d'alguna manera els corpus lingüístics que es descriuen en l'apartat següent a fi de fer-los més útils en l'entrenament

¹³ <http://translate.google.com>

¹⁴ <https://www.bing.com/translator>

¹⁵ <https://www.deepl.com/translator>

¹⁶ S'ha de dir que la recerca sobre TABC per a llengües amb pocs recursos és un camp molt actiu; per exemple, l'autor és el coordinador científic del projecte europeu GoURMET ("Global Under-Resourced Media Translation", <http://gourmet-project.eu>), el qual aborda la traducció automàtica neural entre l'anglès i llengües com l'amhàric, el kirguís o el suahili.

de sistemes basats en corpus; per exemple, indicant a quina categoria lèxica (nom, adjectiu, etc.) pertany cada mot dels textos.

2.3.2 Corpus

En el cas de la traducció automàtica basada en corpus, necessitem sobretot grans quantitats (per exemple, centenars de milers o milions) de parells d'oracions: cada oració amb la seua traducció.¹⁷ La recollida d'aquest tipus de corpus comporta també un esforç considerable. Ha d'existir una quantitat suficient de text traduït per professionals, ha de ser disponible a l'hora d'entrenar el sistema, i les traduccions han d'estar alineades oració a oració (certament, l'alineació de les oracions dels documents i les seues traduccions es pot fer automàticament amb un cert marge d'error). No és estrany que els corpus continguin *soroll*, és a dir, contingut que no es pot considerar una traducció adequada, i que s'ha de detectar i eliminar.

Recentment s'han produït avanços en la recollida automàtica de corpus a partir de webs multilingües (de fet, un dels mètodes que usen els sistemes comercials com Google, Microsoft o DeepL): primerament, es descarreguen documents en les llengües d'interès, s'hi cerca quins documents d'una llengua poden ser traducció dels documents d'altra (examinant-ne la grandària i l'estructura i usant els recursos bilingües disponibles); després se segmenten en oracions, es prova d'emparellar cada oració d'un document amb la seua traducció en l'altre, i, finalment, s'usen tècniques senzilles per a descartar parells d'oracions que no són traducció mútua (per exemple, perquè una és molt més llarga que l'altra).¹⁸

Hi ha projectes com ara OPUS,¹⁹ que intenten arreplegar tots els corpus paral·lels disponibles públicament per a desenes de llengües del món. Tot i això, és comú que per a llengües menors només s'hi dispose de textos religiosos o informàtics.²⁰

2.4 Drets d'ús i llicències

2.4.1 Programari

En el cas dels sistemes de traducció automàtica basada en regles, s'usen dos tipus de programari. D'una banda, com s'ha explicat més amunt, el *motor* que fa la traducció automàtica, el qual hauria de ser tan independent de les llengües com fora possible, i d'altra, les *eines* necessàries per a gestionar els recursos lingüístics que usa el sistema; és a dir, per a editar-los (crear-los des de zero, o actualitzar-los) i convertir-los als formats que usa el motor. Quant a l'accés, el programari de traducció automàtica pot estar pensat per a ser instal·lat en un ordinador local (un ordinador de sobretaula, un portàtil, un telèfon intel·ligent, o un servidor en la nostra institució o empresa) o en un ordinador (un servidor) remot accessible per Internet (com és el cas de Google Translate, DeepL, etc.). En aquest últim cas, els drets d'ús són determinats per les condicions d'ús establides per al servei remot. Tot i que és cert que els sistemes d'Internet donen com més va suport a més llengües menors, hi ha moltes llengües per a les quals encara no hi ha traducció automàtica. En l'apartat 5 es descriuen casos de llengües que només tenen un sistema de traducció automàtica, com ara l'aragonés, el bretó o el sami.

Quan el programari s'ha d'usar en un ordinador local, és particularment pertinent considerar-ne la *llicència*. Tot programari es pot classificar com a *lliure* o *no lliure*. El programari lliure²¹ és el que

- pot ser executat lliurement per a qualsevol propòsit,
- pot ser examinat lliurement per veure com funciona i es pot modificar lliurement per adaptar-lo a una nova necessitat o aplicació (per a això, el *codi font*, és a dir, la forma *editable* del programa, escrita

¹⁷ En el cas de la traducció automàtica estadística, també convé tenir quantitats encara més grans d'oracions en la llengua meta (un corpus de text monolingüe) per a assegurar la naturalitat de les traduccions.

¹⁸ Per exemple, Paracrawl (<http://www.paracrawl.eu>) desenvolupa un programari per a la recollida de corpus bilingües d'Internet i, a més, publica els corpus recollits.

¹⁹ <http://opus.nlpl.eu>

²⁰ D'una banda, de l'Alcorà o dels Testimonis de Jehovà, i d'altra, relacionats amb programari lliure/de codi font obert com ara LibreOffice.

²¹ Podeu trobar la definició a <http://www.gnu.org/philosophy/free-sw.html>.

en un llenguatge de programació, ha d'estar disponible —a més de la forma *executable* que se'n deriva—; d'ací el nom alternatiu de *programari de codi [font] obert*),²²

- pot ser redistribuït lliurement a qualsevol persona i
- pot ser millorat i publicat lliurement de manera que tota la comunitat d'usuaris se'n beneficiï (el codi font ha d'estar disponible també per a això).

Si no es compleixen les condicions indicades, el programari no és lliure (tot i ser gratuït, com ho són, per exemple, el navegador Opera, el visor de documents Adobe Acrobat o el sistema de missatgeria WhatsApp). Per exemple, l'ús pot estar restringit a un ús personal o no comercial, el codi font pot no estar disponible, etc.

El motor i les eines d'un sistema de traducció automàtica per a una llengua menor haurien de ser idealment lliures, com és el cas de la plataforma de traducció automàtica Apertium: la llicència lliure permet millores públiques del motor que beneficien tots els usuaris, independentment de les llengües.²³

2.4.2 Dades

Com hem explicat més amunt en l'apartat 2.3, el programari de traducció automàtica és especial perquè depèn fortament de les dades. La traducció automàtica basada en regles depèn de recursos lingüístics com ara diccionaris morfològics, diccionaris bilingües, gramàtiques i fitxers de regles de transferència estructural; la traducció automàtica basada en corpus (estadística o neural) depèn, directament o indirectament, de la disponibilitat de text paral·lel alineat oració a oració.

Com succeeix sovint en el cas de moltes llengües menors, i malgrat els esforços d'iniciatives com OPUS (esmentat en l'apartat 2.3.2), pot ser encara impracticable obtenir i preparar les quantitats de text paral·lel alineat oració a oració (normalment de l'ordre de centenars de milers o milions de paraules) necessàries per a obtenir resultats raonables en TABC estadística o neural. En casos així pot ser més fàcil que parlants experts de la llengua menor adquirisquen les habilitats necessàries per a codificar els seus coneixements en forma de diccionaris i regles per a un sistema de TABR. Els drets d'accés i les condicions d'ús dels sistemes que se'n deriven dependran òbviament de la llicència amb què es facen disponibles els recursos lingüístics creats. Els recursos (regles, diccionaris) poden ser, com en el cas d'Apertium, lliures —en el mateix sentit que el programari pot ser lliure tal com es descriu en l'apartat anterior—; això fa més efectiu l'efecte del sistema de traducció automàtica resultant sobre l'estatus de la llengua menor, ja que la comunitat lingüística pot no només adoptar-lo i usar-lo, sinó també millorar i difondre els recursos lingüístics en què es basa.

3 Reptes

3.1 Actituds tecnofòbiques o luddites

Encara que una llengua menor tinga un conjunt d'activistes lingüístics motivats i ben formats, cal establir una connexió entre aquesta perícia i les habilitats en tecnologies de la informació. I això pot ser difícil; en moltes comunitats lingüístiques es detecta el que hom podria anomenar actituds *tecnòfobes* o *luddites*:²⁴ persones alfabetitzades en la llengua menor i ben formades desconfien de les tecnologies perquè tenen una visió idealitzada de la llengua i de la comunicació humana o perquè aprecien poc els usos no formals o no literaris.²⁵ Qualsevol grup de persones que s'esforcen a construir sistemes de traducció automàtica de

22 L'Open Source Initiative estableix una definició (<http://www.opensource.org/docs/definition.php>) que és aproximadament equivalent quant als propòsits d'aquest article.

23 Una altra possibilitat seria que el motor i les eines no foren lliures sinó que només foren disponibles públicament i amb formats ben documentats per a les dades: podríem encara construir un sistema creant les dades lingüístiques corresponents, però qui l'executara hauria de tenir dret a usar el motor i les eines.

24 Del pseudònim *Ned Ludd*, usat al Regne Unit per activistes que destruïen les màquines en la revolució industrial per por de quedar-se sense treball. A les nostres terres, a mitjans del segle XIX, aquestes tensions prengueren forma en l'anomenat *conflicte de les selfactines*.

25 Potser perquè molts d'aquests professionals del llenguatge acostumen a centrar-se més en fenòmens generalment poc freqüents que són idiosincràtics d'una llengua determinada (les "joies"), i que no solen ser ben tractats pels sistemes de traducció automàtica, en comptes de fixar-se en com aquests sistemes tracten els mots i estructures comuns que constitueixen el 95% dels textos quotidians (els "maons" o blocs bàsics que construeixen el 95% de la llengua).

codi obert per a un idioma menor han d'estar preparats per a afrontar aquest tipus d'adversitats diguem-ne *socioacadèmiques*.

3.2 Estandardització

La manca d'un sistema d'escriptura comunament acceptat, d'unes normes ortogràfiques o d'un dialecte de prestigi de referència pot suposar realment un desafiament greu per a qualsevol que intente construir un sistema de traducció automàtica per a aquesta llengua menor (es podria anomenar “la síndrome del pioner”): si s'elegeix una norma que finalment no és acceptada majoritàriament, la utilitat del sistema queda compromesa. En l'apartat 5 es discuteixen els casos de l'occità i l'aragonés, llengües amb una norma encara inestable.

Cal puntualitzar que, en el cas de sistemes de traducció automàtica basada en corpus, és possible una aproximació més informal a aquest problema, la qual té clarament conseqüències: si s'entrena un sistema amb corpus amb inconsistències normatives,²⁶ les traduccions automàtiques resultants també en tindran.²⁷

3.3 Elicitació del coneixement i complexitat lingüística

En el cas particular en què calga crear recursos lingüístics, hi ha dos reptes íntimament relacionats.

D'una banda, és possible que per a una llengua menor no hi haja el coneixement lingüístic explícitament codificat que és necessari per a poder crear recursos lingüístics. Per a generar dades lingüístiques útils, el coneixement intuïtiu de la llengua per part dels parlants s'ha de fer explícit, és a dir, s'ha d'*elicitar*. Certament, hi ha tipus de recursos lingüístics que són més senzills de construir que d'altres. Per exemple, amb una interfície de formularis ben dissenyada es pot aprofitar el coneixement lingüístic dels voluntaris per a crear i mantenir diccionaris: per exemple, es pot demanar als voluntaris que introduïren entrades de diccionari monolingües i bilingües mitjançant una interfície de formulari que els permetia seleccionar paradigmes d'inflexió, escollir els equivalents de traducció en qualsevol direcció, etc. Tanmateix, el disseny de determinades porcions de dades lingüístiques necessàries, com ara les regles que transformen les estructures gramaticals d'una llengua en estructures de l'altra, no es presta tan fàcilment al treball de persones no expertes.

D'altra banda, com que caldrà editar i actualitzar els recursos lingüístics, s'ha d'intentar mantenir al mínim el nivell necessari de coneixements lingüístics. L'objectiu és codificar el coneixement de la comunitat lingüística de la llengua mitjançant nivells de representació que es puguin aprendre fàcilment partint d'habilitats i conceptes bàsics de gramàtica i llengua com els que es poden adquirir a l'escola primària o secundària. Això no evita, però, l'aprenentatge dels formats en què aquest coneixement s'ha de codificar; els desenvolupadors novells s'haurien de poder beneficiar de la combinació de documentació escrita i del suport d'altres desenvolupadors.

3.4 Organització de la creació dels recursos

Una de les possibles maneres en què la tecnologia de traducció automàtica de codi obert podria beneficiar un idioma menor —a través de la creació de traducció automàtica que la connecte amb una altra llengua— és el treball de comunitats de persones que desenvolupen recursos lingüístics o recol·lecten i processen textos per a formar corpus. Aquestes persones podrien haver de fer treball voluntari si no hi ha suficient finançament —com sol passar sovint en el cas de llengües menors.²⁸ Molts idiomes menors allunyats de la normalitat o l'oficialitat tenen grups d'activistes, generalment en l'àmbit educatiu, que inclouen persones amb les habilitats lingüístiques i de traducció que els permetrien col·laborar en la creació de dades lingüístiques

²⁶ Certament, algunes inconsistències es podrien eliminar automàticament abans de l'entrenament, però amb conseqüències. Per exemple, la BBC publica textos en Igbo (llengua tonal parlada a Nigèria amb uns 45 milions de parlants), alguns amb marques diacrítics de to i altres sense aquestes marques; com que restituir els diacrítics inequívocament és impossible (precisament pel seu valor diacrític), es pot optar per eliminar-los a tot el corpus, a costa d'introduir ambigüitats.

²⁷ De fet, això s'observa a voltes en traductors com Google Translate.

²⁸ En el cas de la plataforma lliure de traducció automàtica Apertium, es va disposar, inicialment, de finançament públic (Govern d'Espanya, Generalitat de Catalunya) que va permetre contractar personal expert (per als parells espanyol-català, espanyol-gallec, català-aranés, català-anglès, etc.); posteriorment, hi ha hagut parells de llengües que s'han desenvolupat de manera essencialment voluntària, o altres han rebut finançament parcial a través d'empreses, ONGs, o programes com el Google Code-In i el Google Summer of Code.

(diccionaris i regles) i en la creació i gestió de corpus. Però les habilitats lingüístiques i de traducció i el temps de voluntariat no són suficients, encara que siguin completament crucials en el cas dels idiomes menors: el treball dels experts ha de ser coordinat per un grup més reduït de persones que dominen els detalls del motor de traducció i les eines utilitzades.²⁹

3.5 Gestió dels recursos: llicències i comuns

Una bona manera de millorar la situació de les llengües menors, és a dir, amb pocs recursos és crear un cos compartit de recursos lingüístics i programari que siga fàcilment disponible, pugui ser usat lliurement per qualsevol persona i per a qualsevol aplicació, es pugui modificar i millorar fàcilment per a aplicacions noves i anime els desenvolupadors a contribuir-hi amb modificacions i millores. Les llicències lliures (vegeu l'apartat 2.4.1) garanteixen la llibertat en aquestes pràctiques; un factor molt important per a assolir aquests objectius és l'elecció de la llicència.

Fixem-nos que, en contrast, la traducció automàtica comercial dona moltes menys oportunitats. Les principals empreses de traducció automàtica tenen com a objectiu els idiomes del món amb més recursos, ja que s'usen en mercats més desenvolupats on poden fer millor negoci. Això comporta avantatges molt limitats per a les llengües menors. A més, els motors de traducció comercials i els recursos lingüístics que usen són normalment tancats i difícils de modificar per a adaptar-los als idiomes menors sense recursos. Els drets d'ús són molt restringits (per exemple, solen dissuadir o prohibir la barreja i la redistribució) i les llicències, cares, cosa que en dificulta l'adopció. Això comporta que hom hagi d'adreçar-se al venedor per a fer funcionar el sistema per a una nova aplicació. Com a resultat, fa la comunitat lingüística dependent d'un venedor específic.

Després d'haver escollit una llicència lliure per a compartir les dades i el programari, pot passar que, com que poden ser utilitzats, modificats i distribuïts lliurement per qualsevol persona i per a qualsevol aplicació, s'hi acosten desenvolupadors que decidisquen "pujar-hi debades" i usar les dades i el programari per a produir i distribuir productes no lliures sense aportar-hi res a canvi. Idealment, una bona llicència podria intentar dissuadir d'aquesta apropiació privada i de l'aprofitament sense aportar-hi res, tot afavorint el desenvolupament col·laboratiu i l'agregació complementària d'habilitats, sense deixar de permetre negocis al voltant de les llengües implicades (negocis que necessàriament s'haurien de basar en la prestació de serveis més que en la venda de llicències d'ús).

Una possibilitat és elegir llicències que faciliten la creació d'un *comú*. En la vida *analògica*, un comú és un tros de terra d'ús comunitari, és a dir, no dividit, destinat, per exemple, a la pastura, o una zona oberta al públic en un municipi. Per analogia, en la vida digital, podem tenir, d'una banda, *comuns de programari*, és a dir, de codi subjecte a un ús comú i, d'altra, *comuns de recursos lingüístics*. Idealment, els recursos lingüístics i el programari per a la llengua menor haurien de poder ser gestionats com un comú.

El *copyleft* (un joc de paraules sobre el terme *copyright*, drets d'autor, però no obstant això encara un tipus de *copyright*), quan s'afegeix a una llicència lliure, implica que les modificacions s'han de distribuir amb la mateixa llicència. Hi ha, per tant, llicències lliures *sense copyleft* (com ara la llicència BSD de tres clàusules,³⁰ la llicència MIT,³¹ la llicència Apache³² o la llicència Creative Commons Reconeixement CC-BY³³) i *amb copyleft* (com ara la llicència general pública GPL de GNU³⁴ o la llicència Creative Commons Reconeixement Compartir-Igual, CC-BY-SA³⁵).

El *copyleft* dona un suport sòlid a la creació i el manteniment d'un comú de programari o dades, ja que dissuadeix de l'apropriació privada en obligar que qualsevol producte que se'n derive s'ha de distribuir sempre sota els mateixos tèmens i estableix un terreny de joc igual per a tots. Com a resultat, fomenta el

29 El projecte Apertium té un Comitè de Gestió del Projecte (*Project Management Committee*) de set desenvolupadors, que és elegit cada dos anys pels desenvolupadors actius.

30 <https://opensource.org/licenses/BSD-3-Clause>

31 <https://opensource.org/licenses/MIT>

32 <https://www.apache.org/licenses/LICENSE-2.0.html>

33 <https://creativecommons.org/licenses/by/4.0/deed.ca>

34 <https://www.gnu.org/licenses/gpl-3.0.ca.html>

35 <https://creativecommons.org/licenses/by-sa/4.0/deed.ca>

desenvolupament col·laboratiu i permet a les comunitats de desenvolupadors construir conjunts compartits de recursos lliures perquè exigeix que tot treball derivat siga sempre distribuït sota la mateixa llicència lliure. El *copyleft* fomenta negocis basats en serveis (adaptació, instal·lació, redistribució) alhora que desactiva el bloqueig comercial de la llengua menor.

Aquesta secció no pot acabar sense mencionar un problema que afecta els sistemes de traducció automàtica basats en corpus. Com s'ha explicat en l'apartat 2.2.3, aquests sistemes s'entrenen amb traduccions existents, moltes de les quals s'han recollit de documents publicats en Internet. D'una banda, d'acord amb la Convenció de Berna, quan una obra no expressa explícitament els tèmens sota els quals es pot reutilitzar, s'entén que l'autor s'ha reservat tots els drets de reproducció,³⁶ però, la traducció amb un sistema entrenat amb text d'una obra i de l'obra traduïda corresponent comporta realment una reproducció pública de parts substancials l'obra? D'altra banda, és pertinent preguntar-se fins a quin punt es respecten els drets dels autors dels textos originals i de les traduccions quan es reutilitzen per a crear sistemes de traducció automàtica comercials. Si qui ha pagat per la traducció ha deixat clar que es publicaria de manera oberta, no s'ha completat satisfactòriament la transacció amb qui ha traduït? Per a una discussió d'aquests aspectes, vegeu per exemple Moorkens i Lewis (2019).

4 Efectes

L'existència de sistemes de traducció automàtica per a la llengua menor pot tenir efectes positius sobre aquesta. Els efectes seran més intensos com més fàcil siga l'accés als sistemes, i també dependran de l'accessibilitat de les dades usades per a crear-los.

4.1 Normalitat i visibilitat

La disponibilitat de traducció automàtica des d'una de les llengües dominants circumdants pot contribuir a l'augment de la "normalitat" de la llengua menor en el sentit d'estendre'n l'ús familiar i domèstic a contextos socials més formals com ara l'escola, els mitjans de comunicació, l'Administració, el comerç, etc. Només per esmentar alguns exemples:

- Els materials educatius en una de les llengües dominants es poden traduir a la llengua menor per tal que els xiquets estiguen escolaritzats en aquesta llengua.
- Les notícies publicades en una de les llengües majors es poden traduir a la llengua menor per a crear mitjans escrits per a aquesta comunitat lingüística.
- Les lleis, regulacions, informacions governamentals, anuncis, convocatòries, etc. es poden traduir a la llengua menor.
- Les empreses tindrien molt més fàcil comercialitzar nous productes en la llengua menor ("localització"), especialment aquells en els quals el component de text és important com ara l'electrònica de consum, telèfons mòbils, etc.

Per descomptat, en tots aquests escenaris se suposa que és factible publicar el resultat de posteditar els resultats de la traducció automàtica i generar textos adequats. Per tant, els efectes positius esmentats seran més intensos com millor siga el sistema de traducció automàtica; per exemple, quan les divergències lingüístiques entre les llengües implicades siguen menors.

La disponibilitat de la traducció automàtica de la llengua menor a una o més de les llengües principals circumdants pot ajudar a la difusió del material escrit originalment en la llengua menor. Per exemple, el contingut dels llocs web podria ser escrit i gestionat directament en l'idioma menor i traduït automàticament per als usuaris d'altres idiomes principals, ja siga sobre la marxa (en aplicacions d'assimilació com les descrites en l'apartat 2.2.2) o després de ser revisats per professionals (vegeu, per exemple, el cas del sami, que es descriu en l'apartat 5.5).

³⁶ El Conveni de Berna per a la protecció d'obres literàries i artístiques (<https://www.wipo.int/treaties/en/ip/berne/index.html>) estableix que la protecció dels drets d'autor no pot estar condicionada a cap formalitat ("principi de protecció automàtica").

4.2 Alfabetització

La disponibilitat creixent de text en la llengua menor, obtinguda mitjançant la traducció automàtica, la postedició i l'elaboració posterior de material escrit originalment en una llengua dominant, pot motivar els esforços per millorar els nivells d'alfabetització dels parlants d'aquesta comunitat lingüística.

4.3 Estandardització

L'ús de sistemes de traducció automàtica pot contribuir a l'estandardització d'una llengua, per exemple, promovent un sistema d'escriptura particular, una ortografia particular o un dialecte particular (vegeu els casos de l'aragonés i l'occità en la secció 5).

4.4 Increment de l'expertesa i dels recursos disponibles

La creació d'un sistema de traducció automàtica per a la llengua menor implica, fins a cert punt, un procés de reflexió sobre la llengua i comporta una posterior fixació i codificació de coneixements monolingües i bilingües. L'expertesa lingüística resultant, en un entorn de codi obert, resta a disposició de tota la comunitat lingüística amb la publicació dels recursos lingüístics generats.

5 Estudi de casos

En aquest apartat es descriuen breument sis casos on s'ha aconseguit crear un sistema de traducció automàtica útil i disponible públicament entre una llengua menor i una llengua *principal*, els reptes a què s'ha hagut d'enfrontar el desenvolupament, els efectes que ha tingut sobre la llengua menor i els recursos que s'han generat. En vista de l'escassetat de corpus bilingües i l'efecte limitat dels sistemes comercials quan n'hi ha, l'èmfasi és sobre sistemes lliures basats en regles, tots ells basats en la plataforma Apertium.

5.1 Bretó

La llengua bretona (en bretó *brezhoneg*) és una llengua cèltica del grup britònic que es parla a l'oest de Bretanya (*Breizh Izel* o "Baixa Bretanya"), en França, i la llengua principal amb la qual té contacte és el francès, única llengua oficial; de fet, el bretó, parlat per unes 200.000 persones, no té pràcticament cap reconeixement legal a França. Per donar alguns indicadors de la seua situació com a llengua menor, indicarem que només un 2% de l'escolarització es fa en bretó, part de la senyalització viària és bilingüe, i hi ha una presència més bé reduïda de la llengua bretona en mitjans de comunicació. La principal organització promotora de la llengua és l'Ofis Publik ar Brezhoneg,³⁷ i la llengua té una forma estàndard ben establerta i comunament acceptada.

Programes com ara Firefox, el servidor Google i alguns programes de Microsoft com ara Office o Skype han estat *localitzats*³⁸ i hi ha una Wikipedia en bretó amb unes 70.000 pàgines. Hi ha poc programari dedicat a la llengua bretona, la majoria gratuït o lliure com ara el traductor automàtic Apertium (que ara analitzarem) i el corrector ortogràfic i gramatical LanguageTool. Aquest programari i serveis com el diccionari en línia Freelang³⁹ es basen en recursos lingüístics com ara analitzadors morfològics, diccionaris monolingües i bilingües. Quant a corpus de text bilingües, en l'actualitat OPUS conté aproximadament 400.000 parells de frases, la majoria de les quals molt especialitzades, de l'àmbit de la informàtica.

El projecte Apertium conté un traductor automàtic lliure del bretó al francès per a l'assimilació, és a dir, per a permetre als lectors francòfons accedir a contingut escrit en bretó.⁴⁰ Aquest sistema de traducció

37 L'Oficina Pública del Bretó (<http://www.brezhoneg.bzh/>), establiment públic de cooperació cultural (EPCC) fundat per l'Estat francès, el Consell Regional de Bretanya, el Consell Regional del País del Loira i els consells departamentals del Finisterre, el Morbihan, les Costes d'Armor, Illa i Vilena i Loira Atlàntica, té la missió de promoure la llengua bretona i desenvolupar-ne l'ús en totes les àrees d'ús d'una llengua.

38 L'anglicisme *localització* (en anglès *localization*) fa referència al procés d'adaptar un producte a un mercat *local* (regional). Entre les característiques d'un mercat local s'hi sol incloure la llengua.

39 <https://www.freelang.com/enligne/breton.php>

40 Els desenvolupadors van decidir deliberadament no treballar en la traducció del francès al bretó, en considerar-ho massa arriscat pel que fa a la situació sociolingüística, ja que moltes persones podrien donar per bona la traducció automàtica al bretó i usar-la de manera inapropiada pensant que és correcta (Jakez, 2019).

automàtica, l'únic existent en el món per al bretó, és el resultat d'una iniciativa de Francis Tyers —un dels desenvolupadors principals d'Apertium— i va ser presentat a maig de 2009. El sistema fou descrit pel mateix Tyers (2010) i és el resultat de l'esforç conjunt de Gwenvael Jéquel i Fulup Jakez de l'Ofis ar Brezhoneg (predecessora de l'Ofis Publik ar Brezhoneg), de l'empresa valenciana Prompsit Language Engineering i de la Universitat d'Alacant, i és basat en la plataforma Apertium. Els diccionaris no es van construir des de zero, ja que hi havia diccionaris lliures per al bretó en Lexilogos.⁴¹ Una versió primitiva del sistema es va usar (Tyers, 2009; Sánchez-Cartagena et al., 2011) per a ampliar les poques dades (preparades pel mateix Tyers (2009) a partir de materials de l'Ofis ar Brezhoneg i alliberades amb una llicència lliure)⁴² de què es disposava aleshores (unes 31.000 oracions traduïdes) i entrenar sistemes de traducció automàtica estadística.

Recentment, el desenvolupament d'aquest sistema de traducció automàtica ha avançat més lentament, però encara és l'únic disponible per al públic en general i amb una llicència lliure. Quant als problemes de desenvolupament, segons que explica Fulup Jakez, “una de les principals dificultats rau en el temps necessari per millorar els diccionaris i establir regles de transferència. D'altra banda, sempre hem confiat en l'ajuda de Fran[cis Tyers] i mai ens hem fet prou autònoms [d'Apertium] per la nostra manca d'habilitats informàtiques. Per això, quan [Francis Tyers] no ens pot dedicar [...] temps com ho va fer al principi [...] l'activitat del projecte minva”; en qualsevol cas, una de les maneres en què continua millorant és detectant quins són els mots bretons desconeguts per al sistema que són més comuns en els textos que s'envien a traduir a la versió instal·lada en el web de l'Ofis,⁴³ i afegint-los periòdicament al sistema. En l'actualitat, la qualitat del francès generat no és adequada per a ser posteditada, però sí que és suficient perquè una persona francòfona pugui fer-se una idea aproximada del significat d'un text bretó.^{44,45}

Quant a la reutilització dels recursos generats en la construcció del traductor automàtic, una part de les dades s'han usat també per a construir el corrector gramatical LanguageTool⁴⁶ per a la llengua bretona.

5.2 Occità

La llengua occitana o llengua d'oc es designa també amb el nom d'algunes de les seues varietats (lemosí, provençal, gascó, etc.). Aquesta llengua romànica, que va gojar d'un gran prestigi durant l'edat mitjana, és ara una llengua francament minoritzada. En absència d'un recompte oficial, s'estima que té entre 100.000 i 800.000 parlants, principalment en territoris del sud de l'Estat francès, però també a la comarca catalana de la Vall d'Aran (*Val d'Aran* en occità) en la Catalunya administrada per l'Estat espanyol i en zones de l'oest d'Itàlia. L'estandardització de l'occità és encara problemàtica i no és exempta de polèmiques, potser en gran part per la divergència dels seus dialectes,⁴⁷ amb una certa preponderància de l'anomenat llenguadocià en l'estàndard anomenat *occitan referenciau* o *occitan larg*. Aquest *occitan referenciau* està codificat amb el que es coneix com a *nòrma clàssica*, però existeix també un sistema ortogràfic alternatiu, anomenat *nòrma mistralenca*.⁴⁸

L'occità és una llengua menor en molts aspectes. Com s'ha dit, no té pràcticament cap reconeixement legal a França i és reconeguda com a *llengua protegida* a Itàlia. Llevat del cas de la Vall d'Aran, on l'Estatut d'autonomia de Catalunya en garanteix l'escolarització en tant que llengua oficial del país, l'escolarització en occità —en les anomenades *calandretes*— és minoritària —unes seixanta escoles— i es podria dir que

41 <https://www.lexilogos.com/>

42 <http://opus.nlpl.eu/OfisPublik-v1.php>

43 La pàgina del traductor automàtic (<http://www.fr.brezhoneg.bzh/42-traducteur-automatique.htm>) és la més visitada del web de l'Ofis (Jakez, 2009) amb més de 60.000 clics per mes.

44 El 2011, el sistema era capaç de traduir el 90% dels mots en textos provinents de la Wikipedia bretona, usant un diccionari bretó de 17.000 entrades i un diccionari bretó-francès de 26.000 entrades, 250 regles de desambiguació i 250 regles de transformació gramatical.

45 Mentre es revisava aquest article, un article que descriu com usar els recursos del sistema de traducció automàtica bretó-francès d'Apertium per millorar els resultats de la traducció automàtica neural (vegeu l'apartat 2.2.3) en aquest cas en el qual els corpus són escassos ha estat acceptat per a ser presentat al congrés de l'Associació Europea per la Traducció Automàtica, EAMT 2020 (Sánchez-Martínez et al., 2020).

46 <http://languagetool.org>

47 Almenys en la percepció de molts parlants.

48 Anomenada així per ser la usada per l'escriptor Frederic Mistral, guardonat amb el premi Nobel.

tolerada per l'Estat francès.⁴⁹ Un altre exemple: només en pocs llocs de l'Occitània administrada per l'Estat francès es pot trobar alguna senyalització bilingüe (en canvi, en la Vall d'Aran és més sistemàtica). No hi ha tampoc una única organització de referència a càrrec de la llengua.

Quant al programari de propòsit general, una bona part del programari lliure (el navegador Firefox, productes de Google, el sistema LibreOffice, etc.) ha estat *localitzat* i la Wikipedia occitana té uns 90.000 articles. Quant al programari per a la llengua occitana, hi ha dos sistemes de traducció automàtica: un de lliure (occità-català/espanyol/francès), basat en Apertium, i un de no lliure que comercialitza⁵⁰ l'empresa Sail Labs.

Com a resultat del desenvolupament dels sistemes de traducció automàtica s'han alliberat nombrosos recursos lingüístics, tant monolingües com bilingües (diccionaris, regles); els que són lliures estan disponibles a través del projecte Apertium. A més, hi ha recursos consultables com ara Freelang o els diccionaris accessibles a través de Lexilogos.⁵¹ Quant a corpus bilingües, OPUS recull aproximadament 400.000 parells de frases, la majoria procedents de la *localització* de programari lliure.

El desenvolupament del primer sistema de traducció automàtica per a l'occità (occità aranès-català) va començar el 2006 (Armentano-Oller i Forcada, 2006) i va rebre el 2007 l'impuls de la Generalitat de Catalunya com a part del projecte "Traducció automàtica de codi obert per al català" (Universitat d'Alacant i Universitat Pompeu Fabra). Posteriorment, la Generalitat de Catalunya va encarregar a una unió temporal d'empreses (Taller Digital de la Universitat d'Alacant i Prompsit Language Engineering) la creació dels traductors oficials entre l'occità (tant l'aranès com l'occità referencial) i l'espanyol i el català, ja que la reforma de l'Estatut d'autonomia havia establert l'occità com a tercera llengua oficial de Catalunya. De fet, en l'actualitat, hi ha una versió dels traductors encara disponible a través del web de la Generalitat de Catalunya.⁵²

A l'hora de decidir quin seria el model d'occità que produiria el sistema oficial, la Generalitat de Catalunya va crear el 2007 una comissió lingüística dirigida per Aitor Carrera i composta sobretot per lingüistes de prestigi, provinents de diverses regions occitanes de l'Estat francès, de les valls occitanes de Piemont (administrativament italianes) i de la Vall d'Aran, que va fixar el model el 2008 (Carrera, 2008) després de reunir-se deu vegades. Si bé la codificació de l'occità no era un terreny verge, hi restaven nombrosos problemes que van haver de ser resolts. El resultat és fonamentalment una codificació basada en el *languedocià*. Els lingüistes que propugnaven el model alternatiu que hem anomenat *mistralenc* no hi van participar.

Només recentment (2018) s'ha començat a desenvolupar un sistema de traducció automàtica basat en Apertium que connecte l'occità amb la llengua amb què té més contacte, el francès, aprofitant una bona part dels recursos desenvolupats i alliberats durant la construcció dels sistemes anteriors per a l'occità. Aquest sistema encara no té una versió estable.⁵³

La disponibilitat de sistemes de traducció automàtica cap a l'occità han facilitat notablement la creació de contingut en occità a partir de contingut en espanyol o en català; per exemple, articles de la Wikipedia.⁵⁴

5.3 Aragonés

L'aragonés, llengua romànica en situació de perill que es parla en el Pirineu aragonés —sobretot a les valls d'Echo, Ansó i Chistàu, a la zona de Pandicosa i també a la Ribagorça Occidental— té uns 10.000 parlants. Hi ha propostes divergents quant a l'ortografia, cosa que en complica l'ús normal.

49 Les *calandretes* són escoles privades (associatives) que ofereixen ensenyament en immersió lingüística. El sistema públic francès ofereix educació bilingüe en molt poques escoles infantils i primàries, i una certa continuïtat, amb l'occità com a assignatura optativa, en l'educació secundària.

50 O comercialitzava fins fa poc.

51 https://www.lexilogos.com/occitan_dictionnaire.htm

52 <http://traductor.gencat.cat/text.do>

53 El desenvolupament es pot seguir en el repositori <https://github.com/apertium/apertium-oci-fra>.

54 El sistema està integrat dins del servei de traducció d'articles de Wikipedia, "Mediawiki Content Translation": https://www.mediawiki.org/wiki/Content_translation.

La iniciativa de construir el primer traductor automàtic aragonés-espanyol va ser de Juan Pablo Martínez Cortés, professor de Teoria del Senyal i Comunicacions de la Universitat de Saragossa. El 2009 va contactar amb la comunitat d'Apertium. Immediatament, Jim O'Regan, un dels desenvolupadors del projecte Apertium, va crear, usant les dades existents d'espanyol i els paradigmes de flexió aragonesa de Wiktionary, un traductor inicial que Juan Pablo Martínez ha anat ampliant des d'aleshores amb el suport d'altres desenvolupadors com Francis Tyers. Recentment, també s'hi ha afegit un sistema aragonés-català, tancant-se el cercle de les llengües pròpies d'Aragó.⁵⁵ La motivació de Martínez era, en les seues paraules (2019), “fer un traductor automàtic, sobretot per ajudar a les traduccions *no literàries* [...], fonamental en una llengua amb tan pocs recursos, reduint el temps de traducció des del castellà, [...], per al seu ús per l'Administració facilitant [a l'aragonés] un estatus *quasioficial*, [...] o per ajudar i donar seguretat a la gent que aprèn l'aragonès”.

L'ortografia elegida finalment per a l'aragonés és la proposada per l'Estudio de Filología Aragonesa (2010), també usada per moltes iniciatives actives com ara la Wikipedia aragonesa o Softaragonés (que desenvolupa programari per a la llengua aragonesa i distribueix versions traduïdes a l'aragonés de programari lliure, iniciatives de les quals Juan Pablo Martínez era també un dels fundadors). No obstant això, quan el sistema de traducció tradueix de l'aragonés, accepta variants lèxiques, ortogràfiques i morfològiques que no genera quan tradueix a l'aragonés. En l'actualitat, el seu diccionari bilingüe té més de 20.000 entrades i centenars de regles i cobreix el 90% del text de la Wikipedia aragonesa, i és molt usat per a la *localització* del programari a l'aragonés.

En paraules del mateix Martínez, “l'existència del traductor ha donat visibilitat a l'aragonès en diferents camps, i ha ajudat que altres projectes s'enlaïressin (com ara Softaragonés), multiplicant l'efecte de visibilitat externa”. De fet, “en una enquesta feta al 2014 per Internet a 228 persones que parlaven o estudiaven aragonès, un 72% coneixia l'existència del traductor automàtic, i un 41% l'havien fet servir”.

Malgrat això, el Govern d'Aragó ha optat des de 2017 per seguir i promoure una altra ortografia provisional (més semblant a la proposada pel Consello d'a Fabla Aragonesa) i, abandonada la idea d'un sistema de traducció automàtica basat en corpus per l'escassetesa de corpus bilingües (unes 100.000 oracions paral·leles en OPUS), han contractat una empresa externa per a fer una versió d'Apertium en aquesta ortografia.

El cas de la traducció automàtica lliure per a l'aragonés, il·lustra, d'una banda, els beneficis de connectar la llengua menor a *comuns* actius de continguts, programari i recursos lingüístics com ara Apertium, SoftAragonés o Wikipedia, per a produir eines que possibiliten i promouen l'ús de la llengua, però també els riscos de fragmentació normativa que comporta l'ús de les eines lliures per parts d'actors que promouen codificacions alternatives d'una llengua menor que encara no en té una de suficientment estable.

5.4 Traducció entre noruec *bokmål* i noruec *nynorsk*

L'idioma germànic escandinau que anomenem informalment *noruec* es pot considerar un continu dialectal que té dues normes escrites: el noruec *bokmål* (literalment “llengua dels llibres”) és molt similar al danès (com a conseqüència de la creació i ús per part de les elits d'una *koiné* dano-noruega durant el temps de vinculació política amb Dinamarca); el noruec *nynorsk* (literalment “nou noruec”) intenta cobrir les variants basades en el noruec original, no tan influïdes pel danès. Les reformes que s'hi han fet (una de les quals provava de crear un noruec únic o *sammorsk*) han acostat molt els dos estàndards. Com a primera aproximació, es podria dir que el *bokmål* i el *nynorsk* tenen un nivell de divergència (i processos d'interferència i diglòssia) similar al de l'espanyol i català. Noruega es divideix oficialment en zones de domini lingüístic *bokmål* i zones de domini lingüístic *nynorsk*; això afecta, per exemple, l'escolarització —els escolars d'una zona han d'estudiar també l'estàndard de l'altra—,⁵⁶ però la major part dels mitjans i dels continguts són en *bokmål*. Això fa del *nynorsk* una llengua menor.

La proximitat entre el *bokmål* i el *nynorsk* fa que la traducció automàtica siga abordable. De fet, hi ha en l'actualitat dos sistemes. D'una banda, un sistema comercial, i per tant, no lliure, anomenat Nyno⁵⁷

55 <https://softaragones.org/traductor/index.arg.html>

56 Això fa que els estudiants de les zones *bokmål* siguin un dels grups d'usuaris més importants dels sistemes de traducció automàtica al *nynorsk*, ja que —com la premsa noruega ha recollit diverses vegades— els usen com a *ajuda* per a fer els deures.

57 <https://www.nynodata.no/nyno>. La pàgina web (en noruec) indica que, en vista de les divergències quant a l'estandardització, el

tradueix del *bokmål* al *nynorsk* (en l'altra direcció no, potser perquè no hi ha tanta demanda); d'altra, hi ha un sistema lliure basat en Apertium que és bidireccional.⁵⁸ A banda, el fet que el segon es pot usar de *nynorsk* a *bokmål*, es podria dir que la qualitat i la utilitat dels dos sistemes és comparable; la principal diferència és el desenvolupament comunitari en obert de les dades lingüístiques lliures del sistema basat en Apertium. Mentre que els dos sistemes poden facilitar la creació de contingut en la llengua menor,⁵⁹ en aquest cas el *nynorsk*, només el segon ha reutilitzat recursos lliures existents i ha contribuït a la creació de recursos lingüístics lliures (diccionaris bilingües i monolingües, regles de desambiguació, regles de transformació gramatical, etc.). Per això, em centraré principalment en el sistema basat en Apertium.

El desenvolupament del sistema Apertium entre el *nynorsk* i el *bokmål* va començar a març del 2008, i va ser possible gràcies a la disponibilitat de dos recursos lingüístics lliures (ambdós amb la llicència pública general de GNU, que té *copyleft*, vegeu l'apartat 3.5): el Norsk Ordbank (“Banc noruec de mots”) i les regles de desambiguació de l'Oslo–Bergen Tagger, que van ser convertits als formats d'Apertium (Unhammer i Trosterud, 2009). El desenvolupament continua en l'actualitat, malgrat l'existència de qüestions sobre les quals no hi ha acord, com ara si els infinitius dels verbs *nynorsk* han d'acabar en *-e* o en *-a*. En l'actualitat, en el cas de notícies, només cal posteditar un 5% dels mots en les traduccions en brut produïdes pel sistema.

El traductor Apertium *bokmål-nynorsk* és usat massivament pels estudiants, per les agències de notícies NTB i NPK, i per a la creació d'articles de Wikipedia (a través de l'eina de traducció de contingut esmentada quan hem estudiat el cas de l'aragonés). Algunes institucions com el Språkrådet (Consell de la Llengua Noruega) han expressat la seua preocupació pel fet que la traducció automàtica pot estar reduint la varietat del *nynorsk*.

5.5 El traductor de sami del nord a noruec *bokmål*

El sami del nord o sami septentrional és la llengua més parlada del grup sami,⁶⁰ amb uns 20.000 parlants, parlat al nord de Noruega (on és oficial), Suècia i Finlàndia (on és una llengua minoritària reconeguda). Les llengües samis pertanyen al grup ugrofinés i són aglutinants. La llengua té un òrgan normatiu anomenat el Giellagáldu (“Font de la Llengua”).

L'únic sistema de traducció automàtica per al sami septentrional és basat en Apertium, i tradueix només de sami septentrional a noruec *bokmål*, i, per tant, està pensat per a animar a la creació de contingut directament en sami (en xarxes socials, per exemple), contingut que seria accessible als parlants de *bokmål* a través de la traducció automàtica.

El desenvolupament del sistema va començar el 2010, principalment per Trond Trosterud i Lene Antonsen, membres del grup Giellatekno de tecnologies per a les llengües samis de la Universitat Àrtica de Noruega⁶¹ en Tromsø, i el desenvolupador d'Apertium Kevin Unhammer, i es va basar en recursos lingüístics lliures de Giellatekno (analitzadors i generadors morfològics). Es tracta d'un parell de llengües molt diferents, i una bona part del desenvolupament fet (i de l'encara pendent) afecta les regles de transformació d'estructures sintàctiques. Els recursos lingüístics generats poden servir, per exemple, per a crear nous sistemes de traducció automàtica entre el sami i el *nynorsk* o el suec, els dos similars al *bokmål*.

Hi ha poques dades sobre l'ús del sistema Apertium sami septentrional-*bokmål*, però tot apunta a un ús bastant intens si es té en compte el nombre de parlants. El traductor automàtic sami septentrional-*bokmål* de la Universitat Àrtica de Noruega⁶² rep aproximadament un centenar de visites al dia, amb unes 5 accions (traduccions) per visita. Aquesta activitat de traducció automàtica per al sami septentrional se suma a la que té lloc en la web principal d'Apertium, la qual també ofereix el parell de llengües.

sistema pot traduir a tres nivells de *nynorsk*, “radikal, moderat eller konservativ *nynorsk*”.

58 De fet, és un dels sistemes més usats dels basats en Apertium, particularment per estudiants.

59 Per exemple, l'ús de *nyno* va facilitar la *localització* del programari de Microsoft al *nynorsk* (Unhammer, 2009).

60 La denominació *sami* substitueix la tradicional *lapó*, que es considera ara inadequada.

61 <http://giellatekno.uit.no>

62 <http://jorgal.uit.no/index.sme.html?dir=sme-nob>

5.6 El traductor automàtic de la Generalitat Valenciana

El català no és en l'actualitat tan *menor* com les cinc llengües estudiades en aquest apartat, ja que disposa de més recursos i, entre ells, diverses opcions de traducció automàtica, però el cas que discutirem en aquest apartat il·lustra com l'elecció d'un determinat model de desenvolupament i d'accés de les dades pot posar en perill un sistema útil en el qual s'han invertit molts diners públics.

Tot i estar pràcticament acabat l'any 1997, la Generalitat Valenciana no va publicar fins a l'any 2000 SALT, el primer sistema de traducció automàtica⁶³ espanyol-català (valencià) d'accés relativament general.⁶⁴ Inicialment, SALT, desenvolupat per un equip dirigit per Josep Lacreu, es distribuïa gratuïtament en forma de disquets, però només en la seua forma executable, i era concebut com un sistema d'ajuda a la redacció de textos en valencià, amb un cert nivell d'interactivitat en cas d'ambigüitat. Ni les dades lingüístiques que usava —que incloïen excel·lents diccionaris— ni el codi font del programa eren públics. Les versions posteriors (fins a la versió 4) es distribuïen en CD-ROMs o com a paquets descarregables d'Internet i també es van fer disponibles per a usar-les directament en Internet, inicialment amb restriccions.

Les eleccions valencianes del 2015 van propiciar un canvi en el Consell de la Generalitat Valenciana, i els consegüents canvis en el personal que s'encarregava de SALT. En el canvi, el personal ixent va *oblidar* les claus d'accés a les bases de dades on es guardaven les dades lingüístiques i els nous responsables no en van poder continuar el desenvolupament. Afortunadament, hi havia dades lliures molt completes per a la traducció castellà-valencià en Apertium, i no va ser necessari començar de zero:⁶⁵ amb el suport de l'empresa Prompsit Language Engineering i de la comunitat de desenvolupadors d'Apertium, el mateix personal de la Generalitat Valenciana en va fer el desenvolupament. En l'actualitat, SALT és un programari completament nou, accessible per Internet⁶⁶ i disponible com a aplicació per a mòbil Android i iPhone, basat en la plataforma de traducció automàtica lliure Apertium. Les dades lingüístiques, completament lliures, formen part dels recursos lingüístics generals apertium-spa (espanyol),⁶⁷ apertium-cat (català)⁶⁸ i apertium-spa-cat (espanyol-català)⁶⁹ d'Apertium,⁷⁰ on han estat ampliadades i adaptades als usos de la Generalitat Valenciana.⁷¹ La preservació dels recursos usats pel nou SALT queda, per tant, reforçada mitjançant l'ús d'un repositori públic allotjat en un dels principals serveis (GitHub), lluny de la situació de vulnerabilitat que va forçar el canvi tecnològic.

6 Conclusions

Una de les eines informàtiques que pot tenir més impacte en la presència en línia d'una llengua, i, per tant, en la seua vitalitat en un món on com més va més comunicació es produeix de manera digital, és la traducció automàtica. La traducció automàtica, que es pot usar tant per a comprendre contingut en una altra llengua com per a generar-ne, pot usar *recursos lingüístics* com ara diccionaris i regles escrits per persones expertes o *aprendre* a partir de *corpus* amb moltíssims exemples d'oracions traduïdes. En el cas, bastant comú en el cas de llengües *menors*, que no es puga aplegar una col·lecció suficientment extensa d'oracions traduïdes, les comunitats lingüístiques han de trobar la manera de construir, mantenir i publicar els recursos lingüístics necessaris amb llicències que maximitzen l'efecte positiu sobre la llengua en desavantatge. Després d'una breu introducció als usos i les tecnologies de traducció automàtica, s'han descrit els obstacles genèrics a la construcció de sistemes de traducció automàtica per a llengües menors i els efectes que hi poden tenir. Finalment, s'han estudiat en més detall els casos de sis llengües europees *menors*: en els sis casos, els

63 Es podia usar en mode automàtic i en mode interactiu, resolent manualment les ambigüitats que aparegueren durant la traducció del text.

64 Una de les raons del retard va ser la falta d'acord quant al model de valencià que el sistema havia de produir. El nom són les sigles del Servei d'Assessorament Lingüístic i Traducció, que l'impulsava la Generalitat Valenciana.

65 Tot i que es podrien haver obtingut algunes dades lingüístiques per enginyeria inversa de SALT 4.0.

66 <http://www.salt.gva.es/>

67 <https://github.com/apertium/apertium-spa>

68 <http://github.com/apertium/apertium-cat>

69 <http://github.com/apertium/apertium-spa-cat>

70 En els diccionaris, les entrades que divergeixen del model general es marquen amb “val_gva”.

71 <http://www.salt.gva.es/va/criteris>

sistemes construïts s'han basat en recursos lingüístics lliures i s'han muntat sobre la mateixa plataforma lliure de traducció automàtica, Apertium. En alguns casos s'han abordat reptes com ara els relacionats amb l'estandardització; en la majoria d'ells s'han detectat efectes positius com un augment de l'ús *en línia* de les llengües menors i la creació de recursos lingüístics que es poden usar per a altres llengües.

Agraïments: a Juan Pablo Martínez (Universitat de Saragossa), Trond Trosterud (Universitat Àrtica de Noruega, Tromsø), Kevin Brubeck Unhammer (Trigram AS), Robert Escolano (Universitat d'Alacant), Fulup Jakez (Oficina Pública de la Llengua Bretona), Gema Ramírez-Sánchez (Prompsit Language Engineering), Carme Armentano i Oller (traductora i lingüista computacional autònoma a Lausana), i als revisors de l'article pels seus interessants suggeriments.

Referències bibliogràfiques

- Armentano-Oller, Carme, i Forcada, Mikel L. (2006). Open-source machine translation between small languages: Catalan and Aranese Occitan. Dins *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)* (p. 51–54) [organitzat en conjunció amb LREC 2006 (22-28.05.2006)].
- Carrera, Aitor. (2008). [Acòrds dera Comission Lingüística deth traductor automatic catalan-occitan-occitan-catalan](#)". (164 p.).
- Casacuberta Nolla, Francisco, i Peris Abril, Álvaro. (2017). Traducció automàtica neuronal. *Tradumàtica: Tecnologies de la Traducció*, 15, 66–74.
- Forcada, Mikel. (2006). Open source machine translation: an opportunity for minor languages. Dins *Proceedings of the Workshop "Strategies for developing machine translation for minority languages"*, LREC (vol. 6) (p. 1-6).
- Forcada, Mikel L. (2017). Making sense of neural machine translation. *Translation Spaces*, 6(2), 291-309.
- Estudio de Filología Aragonesa. (2010). [Propuesta ortografica provisional de l'Academia de l'Aragonés](#). Saragossa: Edicions Dichitals de l'Academia de l'Aragonés.
- Eurobarometer Special. (2012). [Europeans and their Languages](#). Comissió Europea.
- Jakez, Fulup. (2019). [Comunicació personal].
- Martínez, Juan Pablo. (2019). [Comunicació personal].
- Moorkens, Joss, i Lewis, Dave. (2019). [Research questions and a proposal for the future governance of translation data](#). *Journal of Specialised Translation*, 32, 2-25.
- Nurminen, Mary. (2019). Decision-making, risk, and gist machine translation in the work of patent professionals. Dins *Proceedings of the 8th Workshop on Patent and Scientific Literature Translation* (p. 32-42).
- Rivera Pastor, Rafael, Tarín Quirós, Carlota, Villar García, Juan Pablo, Badia Cardús, Toni, i Melero Nogués, Maite. (2017). [Language equality in the digital age: Towards a human language project](#) [Estudi IP/G/STOA/FWC/2013-001/Lot4/C2]. Parlament Europeu.
- Sánchez-Cartagena, Víctor M., Sánchez-Martínez, Felipe, i Pérez-Ortiz, Juan Antonio. (2011). Enriching a statistical machine translation system trained on small parallel corpora with rule-based bilingual phrases. Dins *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* (p. 90-96).
- Sánchez-Cartagena, Víctor M., Forcada, Mikel L., i Sánchez-Martínez, Felipe. (2020). A multi-source approach for Breton-French hybrid machine translation [acceptat per a ser presentat en el 22nd

Annual Conference of the European Association for Machine Translation (EAMT 2020), del 3 al 5 de novembre de 2020].

- Streiter, Oliver, Scannell, Kevin P., i Stuflessner, Mathias. (2006). Implementing NLP projects for non-central languages: instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4), 267-289.
- Tyers, Francis M. (2009). Rule-based augmentation of training data in Breton-French statistical machine translation. Dins *Proceedings of the 13th Annual Conference of the European Association of Machine Translation EAMT09* (p. 213-218).
- Tyers, Francis M. (2010). Rule-based Breton to French machine translation. Dins *Proceedings of the 14th Annual Conference of the European Association of Machine Translation* (p. 174-181).
- Unhammer, Kevin. (2019). [Comunicació personal].
- Unhammer, Kevin, & Trosterud, Trond. (2009). Reuse of free resources in machine translation between Nynorsk and Bokmål. Dins Pérez-Ortiz, Juan Antonio, Sánchez-Martínez, Felipe, i Tyers, Francis M. (coords.) *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation* (p. 35–42) [congrés celebrat a la Universitat d'Alacant el novembre de 2009].
- Williams, Briony, Nadeu, Climent, Sarasola, Kepa, Ó'Cróinin, Donncha, i Petek, Bojan. (2001). Speech and language technology for minority languages. Dins *Proceedings of Eurospeech 2001*.