REVISTA DE LLENGUA I DRET // JOURNAL OF LANGUAGE AND LAW #73

SOFTCATALÀ: NOUS REPTES PER GARANTIR LA VITALITAT DEL CATALÀ A LES TECNOLOGIES

Marc Riera Irigoyen, Xavier Ivars Ribes, Pere Orga Esteve, Joan Montané Camacho, Jordi Mas Hernández, Artur Vicedo Cremades*

Resum

Aquest article presenta les necessitats actuals en l'àmbit digital per garantir la vitalitat de les llengües minoritzades i sense estat, com el català, en les tecnologies més recents i futures, des del punt de vista de la tasca de Softcatalà amb el català en el programari lliure. En primer lloc, es descriuen una sèrie de tecnologies lingüístiques existents i en desenvolupament de l'associació i el seu funcionament bàsic. A continuació, es parla de la necessitat d'alliberar dades amb formats estàndard i oberts per promoure l'aparició de noves iniciatives i l'aprofitament de recursos. Finalment, es tracta la importància de la legislació sobre la situació de la llengua a la tecnologia.

Paraules clau: Softcatalà, recursos lingüístics, dades obertes, formats oberts, llicències lliures, llengües minoritzades.

SOFTCATALÀ: NEW CHALLENGES TO ENSURE THE VITALITY OF CATALAN IN TECHNOLOGIES

Abstract

This article presents the current needs in the digital environment to ensure the vitality of minority languages without a state of their own, such as Catalan, in the most recent and future technologies, from the viewpoint of Softcatalà's work with Catalan in free software. Firstly, a series of existing linguistic technologies, those being developed by the association, and how they operate are described. We then discuss the need to release data in standard, open formats to promote the appearance of new initiatives and make good use of resources. Finally, we deal with the importance of legislation concerning the situation of language in technology.

Keywords: Softcatalà; linguistic resources; open data; open formats; free licences; minority languages.

*Membres de Softcatalà, associació que des de 1998 promou l'ús del català a les tecnologies i el món digital. info@softcatala.org

Citació recomanada: Riera Irigoyen, Marc, Ivars Ribes, Xavier, Orga Esteve, Pere, Montané Camacho, Joan, Mas Hernández, Jordi, i Vicedo Cremades, Artur. (2020). Softcatalà: nous reptes per garantir la vitalitat del català a les tecnologies. *Revista de Llengua i Dret, Journal of Language and Law*, 73, 146-153. https://doi.org/10.2436/rld.i73.2020.3396

Sumari

1 Introducció

La minorització que pateix el català als territoris de parla catalana no s'ha reproduït de manera tan forta en el món de les noves tecnologies de la informació i la comunicació. Tot i que el català mai no ha format part de les grans llengües d'Internet (anglès, francès, castellà, xinès, alemany), ha demostrat una força molt gran en comparació amb altres llengües semblants respecte al nombre de parlants o del territori on es parla.

Un dels motius que expliquen la fortalesa de la llengua a Internet i en les noves tecnologies és, sens dubte, la implicació dels mateixos parlants de la llengua. Són coneguts els exemples de la Wikipedia en català, la Viquipèdia, primera Wikipedia en llengua no anglesa a tenir contingut, o el Navegador en català, la primera traducció del Netscape Navigator feta des del voluntariat, que va ser la presentació pública de Softcatalà. Softcatalà també va col·laborar amb altres empreses com Google per traduir la interfície del seu cercador.[1]

A més d'aquestes iniciatives nascudes des de la societat civil i el voluntariat, també en van aparèixer d'altres per part d'institucions públiques, com el suport de la Generalitat de Catalunya per traduir parcialment el Windows en català o, per exemple, l'aparició del portal Yahoo! gràcies al suport dels governs català, andorrà i balear («Yahoo!, en català», 2001).

Així, no sense dificultat, des de principis de segle era possible *viure en català* en l'àmbit de les TIC: hi havia sistemes operatius totalment (diverses distribucions de Linux) i parcialment (Windows) traduïts al català, programes ofimàtics (OpenOffice.org i, més endavant, LibreOffice) i navegadors i gestors de correu electrònic (Mozilla i els hereus Firefox/Thunderbird).

Tanmateix, el que s'ha pogut fer i aconseguir durant aquests últims vint anys és només una base sobre la qual cal seguir treballant, sense perdre de vista les novetats que han irromput al mercat (com els assistents de veu) que necessiten una dedicació d'esforços en nous àmbits i de noves maneres.

2 Softcatalà: tecnologies lingüístiques

Tot i que Softcatalà va néixer a partir de la traducció d'un programa (Netscape) i la localització del programari lliure perquè estigui disponible en català, segueix sent l'eix principal de treball, ha anat ampliant l'àrea de treball i millorant la manera de treballar amb l'ajuda de la tecnologia i ofereix, entre altres recursos, diferents tecnologies lingüístiques de suport al català. N'esmentem les més rellevants: per una banda, les dues eines amb més pes amb diferència (corrector i traductor) i, per altra banda, les eines de suport a l'aprenentatge i el projecte Common Voice. Aquest darrer no és un projecte propi, però és clau perquè el català no sigui una llengua de segona en les innovacions tecnològiques més recents.

2.1 Corrector

La correcció de textos és una de les utilitats més bàsiques que ofereixen les tecnologies lingüístiques, i pot ser present en qualsevol dispositiu en què es treballi amb text. El corrector de Softcatalà és el servei amb més ús al web (l'any 2019 es va visitar la pàgina 20 milions de vegades, i es van fer 66 milions de correccions). A més de la correcció purament ortogràfica, també es comproven errors gramaticals i d'estil.

El corrector de Softcatalà, limitat inicialment a la correcció d'errors ortogràfics, es basa des de l'any 2013 en la tecnologia de LanguageTool, un programa lliure de correcció lingüística. Softcatalà contribueix a la millora de les regles per al català, que actualment és una de les llengües més desenvolupades. A més de la pàgina web de Softcatalà, LanguageTool té connectors per a navegadors web, per al LibreOffice i per al Microsoft Word, així com una aplicació per a Android. Gràcies al fet que també es pot utilitzar a través d'una API, es pot integrar fàcilment en altres programes o processos, i a Softcatalà també s'utilitza per cercar errors a les memòries de traducció de programari lliure.

A diferència d'altres tecnologies lingüístiques menys personalitzables, el corrector permet que l'usuari activi o desactivi regles per adaptar-lo a les seves necessitats. El complement per al LibreOffice permet accedir a la llista completa de regles, i la versió web de Softcatalà disposa d'un menú de configuració més reduït

1 Consulteu-ne la notícia.

però amb les opcions més habituals, com l'ús dels diacrítics tradicionals o no, el model de formes (generals, valencianes o balears) i el tipus de cometes, entre d'altres.

2.2 Traductor

El servei de traducció automàtica de Softcatalà ofereix la possibilitat de traduir textos entre el català i set llengües (castellà, anglès, occità/aranès, romanès, francès, portuguès i aragonès). En la direcció castellà>català, és el segon servei de Softcatalà pel que fa a l'ús (amb més de 13 milions de pàgines vistes i 56 milions de traduccions fetes), només per darrere del corrector. És una eina de gran importància, ja que facilita la creació de text en català i es pot aprofitar, per exemple, per accelerar la localització de programari al català.

Inicialment s'utilitzava el traductor Internostrum, desenvolupat pel Departament de Llenguatges i Sistemes Informàtics de la Universitat d'Alacant cap a l'any 2000 (Canals-Marote et al., 2001), i s'oferia un servei de traducció només entre el castellà i el català. A partir de l'any 2011, es va passar a utilitzar Apertium (Ivars-Ribes i Sánchez-Cartagena, 2011), una plataforma lliure de traducció automàtica basada en regles hereva d'Internostrum. Gràcies a aquest canvi, es va poder integrar el traductor al servidor de Softcatalà, es va ampliar el nombre de llengües disponibles i es va millorar la qualitat de les traduccions, perquè es poden incorporar totes les millores en el desenvolupament d'Apertium a la instal·lació al servidor de Softcatalà. A més, les millores creades per part de Softcatalà s'incorporen al codi font general d'Apertium, de manera que tothom pot beneficiar-se'n, independentment de si utilitzen el servei en línia de Softcatalà o instal·len Apertium al seu ordinador.

La plataforma Apertium (Forcada et al., 2011), comprèn una sèrie de programes (anomenats *mòduls*) especialitzats en una part específica de la traducció que processen el text d'origen en cadena fins a generar el text de destinació. Per fer-ho, utilitzen dades lingüístiques codificades com ara diccionaris lèxics o regles gramaticals; la separació entre el codi dels programes (compartits entre parells) i les dades lingüístiques (específiques per a cada llengua i parell de llengües) facilita el desenvolupament i el manteniment dels parells de traducció.

En tractar-se d'un projecte de programari lliure, amb una comunitat rica al voltant, Softcatalà ha participat més enllà de millorar els parells de llengües que s'ofereixen des del nostre lloc web. Mitjançant programes com el Google Summer of Code, membres de Softcatalà han participat en el projecte enfocat a estudiants universitaris des de l'any 2017.[2] Així mateix, ha participat en les proves i desenvolupament de nous mòduls utilitzats per altres parells de llengües.

Els sistemes de traducció automàtica basada en regles, com Apertium, són sovint l'única manera d'aconseguir resultats mitjanament bons per a llengües amb pocs recursos. Malgrat l'enorme esforç necessari per desenvolupar un traductor automàtic amb aquesta tecnologia, ja que el desenvolupador ha de tenir competències lingüístiques de les llengües implicades, la inexistència o escassetat de corpus monolingües i bilingües fa que sigui inviable pensar en tecnologies més recents, com els sistemes estadístics o de xarxes neuronals, que permeten crear traductors automàtics fàcilment sempre que hi hagi prou dades. Les proves que hem fet a Softcatalà entrenant models de traducció automàtica neuronal amb textos bilingües en anglès i català extrets de traduccions existents de programari lliure han estat satisfactòries, però la realitat és que no tenim prou dades textuals de caràcter general amb llicència oberta que ens permetin substituir a curt termini les eines de traducció actuals per unes basades en corpus.

2.3 Suport a l'aprenentatge

Les tecnologies lingüístiques de suport a l'aprenentatge comprenen un conjunt de recursos de consulta orientats a ajudar els usuaris en procés d'aprendre el català o que volen resoldre un dubte lingüístic. Aquests recursos inclouen el diccionari de sinònims i el diccionari multilingüe, així com altres eines de desenvolupament recent.

El diccionari de sinònims, basat en OpenThesaurus, rep un nombre considerable de visites malgrat tenir un desenvolupament limitat. El diccionari es pot consultar a través del web i instal·lar a l'ordinador, en

2 Consulteu-ne la notícia.

aplicacions com el LibreOffice. Softcatalà té la intenció de millorar-lo. Com que el diccionari és lliure, les dades del diccionari podrien servir, a més, per millorar altres eines com el LanguageTool. Pel que fa a la resta de recursos de suport a l'aprenentatge, trobem un separador per síl·labes (amb l'opció de fer la separació sil·làbica de versos), una eina per determinar com s'indica l'hora en català, un conjugador verbal i un convertidor de xifres a nombres escrits, entre d'altres.

2.4 Common Voice

Common Voice és un projecte de Mozilla iniciat l'any 2017 que consisteix en la creació d'un conjunt de dades de veu per a diverses llengües. Tot i que no es tracta d'una iniciativa de Softcatalà, ens hem encarregat de traduir el lloc web del projecte, de crear el recull de frases en català que els usuaris han de llegir i de promoure el projecte entre la comunitat catalanoparlant per implicar-hi el màxim de gent possible i fer créixer el nombre d'hores enregistrades. Un dels reptes inicials ha estat aconseguir frases en domini públic, atès que la majoria d'obres en domini públic són en ortografia prefabriana.

L'objectiu de Common Voice és crear un corpus de veu de domini públic (llicència Creative Commons CC0), de manera que qualsevol persona o empresa pugui aprofitar-lo per crear productes que utilitzin aquestes dades. Principalment, està enfocat als sistemes de reconeixement de la parla, que actualment estan desenvolupats per grans empreses tecnològiques amb prou diners i es limiten a un nombre reduït de llengües, en funció dels interessos comercials. Per a llengües amb pocs parlants, com el català, és l'única manera de reunir les dades necessàries per desenvolupar sistemes lliures de reconeixement de la parla.

A finals del 2019, el nombre d'hores validades en català era de 233 de les 1.000 considerades com a mínim de referència per entrenar un model de reconeixement de la parla de qualitat acceptable. Softcatalà té un grup de Telegram dedicat a les tecnologies de la parla centrat actualment en el projecte, i la iniciativa ha tingut ressò en alguns mitjans de comunicació catalanoparlants i a les xarxes socials.

Cal tenir present que sense un corpus de veu en català no és possible entrenar cap motor de reconeixement de veu en català. I sense això no és possible desenvolupar un assistent de veu com Siri o Alexa. Així doncs, Common Voice és un primer pas necessari del camí en la implementació del processament del llenguatge natural en català.

3 Llicències i formats oberts: una necessitat clau

Per incentivar l'ús del català en les noves tecnologies cal que les institucions públiques o subvencionades que produeixen continguts de referència en català publiquin les seves obres en una llicència que en permeti la lliure distribució. És tant o més important que els continguts es distribueixin en formats estàndards i oberts per assegurar la interoperabilitat en diferents sistemes i dispositius al llarg del temps. En la majoria de casos, és beneficiós també que les llicències siguin lliures: que permetin l'ús comercial i la creació d'obres derivades. Exemples d'aquestes llicències són la GPL per al programari i la Creative Commons BY-SA per a continguts. Addicionalment, en alguns casos, es tendeix a escollir una llicència poc restrictiva per promocionar-ne l'ús. Aquest és el cas per exemple dels projectes Wikidata i Common Voice, les dades dels quals es publiquen en la llicència de domini públic Creative Commons 0.

El problema de la publicació d'obres privatives i en formats propietaris és especialment lesiu per a les comunitats de parlants de llengües minoritàries o minoritzades com el català, on les possibilitats que la comunitat produeixi continguts lliures de qualitat són menors a causa del nombre reduït de parlants. A més, si aquestes obres privatives són creades per organismes públics, representen un malbaratament de recursos i esforços, atès que es perden oportunitats per a la llengua. Aquest malbaratament no es limita només a la llengua i és més evident en el cas del programari creat per organismes públics, que es pot aprofitar en altres projectes si se'n publica el codi font. En aquest sentit, la campanya Public Money, Public Code, de la qual Softcatalà va signar la carta oberta, intenta que la ciutadania conegui aquesta problemàtica i reclami l'alliberament de programari amb llicències lliures per part de les institucions.

Més enllà dels recursos lingüístics bàsics com els que ofereixen Softcatalà (corrector, traductor, etc.) o les entitats i institucions (diccionaris i altres obres de referència), que són una base absolutament necessària,

l'evolució de la tecnologia cap a l'aprofitament de dades massives per crear sistemes d'intel·ligència artificial fa que sigui indispensable l'alliberament de dades de manera oberta per estimular i garantir la vitalitat de la llengua, especialment les llengües minoritzades i sense estat com el català. Sense aquestes dades, resulta impossible desenvolupar, per exemple, sistemes de traducció automàtica d'última generació o sistemes de reconeixement de veu, que progressivament han guanyat protagonisme i es poden trobar en un grup reduït de llengües amb un gran pes comercial, però que no és així per a les llengües més petites i sense un estat que les defensi.

Un altre exemple de la problemàtica dels continguts privatius és el cas de l'estat del català en els lectors de llibres electrònics (*e-readers* en anglès). Actualment, no hi ha cap lector de llibres que incorpori de sèrie un diccionari de llengua catalana integrat en la interfície (cosa que permet consultar les definicions còmodament durant la lectura). Es troben a la venda alguns diccionaris en català per a dispositius Amazon Kindle, però no és possible instal·lar-los en cap altre dispositiu: no es distribueixen en una llicència que ho permeti i només es poden descarregar en el format propietari d'Amazon, que no és compatible amb altres fabricants. El *Diccionari de la llengua catalana* de l'Institut d'Estudis Catalans, per exemple, no es troba disponible per instal·lar legalment en cap lector de llibres electrònics.

El *Diccionari de sinònims* d'Albert Jané ha estat publicat per l'IEC en els últims anys en una versió en línia amb la llicència Creative Commons BY-NC-ND. Tot i que la llicència escollida és molt restrictiva (es permet la còpia i distribució en qualsevol mitjà i format, però no l'ús comercial o la producció d'obres derivades), ha permès a la comunitat convertir-lo a diferents formats, cosa que ha fet possible instal·lar-lo i utilitzar-lo fora de línia en diversos aparells i aplicacions, un ús que (segurament) en un inici no s'havia previst.

4 Legislació i llengua en la tecnologia

Quan les empreses decideixen afegir una llengua en els seus productes, ho fan, principalment, per dos motius: per interessos comercials i per complir la legislació vigent.

Tenint en compte el paper que juga l'anglès en el món actual, i que una part molt important dels productes tecnològics es crea en països angloparlants, no resulta estrany que el primer pas per a qualsevol empresa sigui crear un producte tecnològic en aquesta llengua. D'aquesta manera, per exemple, un fabricant de mòbils farà que el seu assistent de veu funcioni en anglès, i després anirà afegint noves llengües en funció del mercat que vulgui cobrir. El mateix passa amb tecnologies no tan avançades, com pot ser la llengua dels menús, els teclats o els diccionaris; els interessos comercials són suficients perquè les empreses considerin localitzar els seus productes a les llengües dominants i de major pes comercial.

En els casos en què una llengua no es considera rellevant, els estats poden regular i obligar les empreses a afegir el suport d'una llengua determinada, o a fer-ho a través d'algun estàndard determinat. Per exemple, l'any 1985 l'Estat espanyol va regular amb el Reial Decret 2707/1985 com havia de ser la disposició dels teclats dels aparells electrònics comercialitzats a l'Estat, i va obligar a incloure-hi la lletra enya (*ñ*). A partir de llavors, tots els teclats físics venuts a l'Estat espanyol van complir amb la regulació i ara, tot i que el Reial decret que hi obligava ja no és vigent, es continua complint, perquè els fabricants s'han adaptat al mercat. En un exemple més proper en el temps, l'any 2016 l'Índia va determinar que tots els mòbils havien de permetre la introducció de text en anglès, hindi i almenys una altra de les llengües oficials de l'Estat, i permetre mostrar text en les 22 llengües oficials. Després d'algunes pròrrogues, l'any 2018 va entrar en vigor aquesta obligació, coincidint amb l'increment de llengües asiàtiques admeses a Android i iOS. Per tant, les competències d'un estat són útils per garantir el suport tecnològic d'una llengua.

En el cas de les llengües minoritzades i sense estat, com passa amb el català, l'absència d'un ens públic amb competències suficients per vetllar per la plena incorporació de la llengua pròpia en els productes tecnològics dificulta la normalització lingüística en l'àmbit tecnològic. Així doncs, els organismes públics només poden exercir pressió com a client potencial, i la incorporació d'aquestes llengües depèn íntegrament dels esforços de la societat civil (per exemple, des del programari lliure).

5 Les llengües minoritàries en la tecnologia més enllà del 2020

El poder que tenen els grans fabricants ara és superior al que tenien als anys noranta: mentre que fa vint anys tot es feia en ordinadors sobre els quals l'usuari podia tenir tot el control, cada vegada més tot passa per mòbils de capacitats més limitades i, especialment, el núvol. El salt als servidors dels fabricants i les empreses tecnològiques suposarà que la capacitat d'acció de particulars i associacions serà més reduïda i seran les grans corporacions les que decidiran quines llengües tenen accés a quins serveis.

Fins ara, el text era el suport principal. La informació en format text tendeix a tenir un suport universal; els fabricants han decidit admetre el màxim nombre de llengües possible. Això és un repte tecnològic que, via el consorci Unicode, any rere any, augmenta el nombre de llengües que poden codificar text digitalment. Els fabricants, al seu torn, afegeixen els teclats i lletres digitals per poder introduir i renderitzar els textos.

Últimament, amb l'aparició de dispositius intel·ligents (sovint sense interfície gràfica) connectats a la xarxa, ha pres protagonisme la tecnologia de veu. La veu és completament diferent del text: el nombre de llengües admeses és molt inferior i, tot i que els fabricants han anat afegint llengües, no són tan ambiciosos com amb el text. A això cal sumar-hi el principal problema: el sector és un oligopoli on els fabricants tenen el control total sobre l'ecosistema de sistema operatiu i aplicacions, de manera que si no s'admet una llengua, no hi ha res a fer. Es crea, doncs, una línia vermella on algunes llengües quedaran per sota i excloses dels productes i serveis relacionats amb la veu. L'efecte és pervers, perquè per gaudir d'aquests serveis els parlants de les llengües no admeses es veuen obligats a configurar el seu aparell en una llengua dominant, i això fa que la llengua minoritzada desaparegui completament de les estadístiques i dificulta encara més que les empreses considerin afegir suport per a les llengües minoritzades.

Tenint en compte aquesta nova realitat, les iniciatives que sorgeixin de la comunitat tindran encara més importància per determinar si els nous productes tecnològics basats en la veu estan disponibles en les llengües més petites amb poc o gens d'interès comercial i recursos limitats. Cal aprofitar al màxim els recursos existents alliberant-los amb llicències obertes que en promoguin l'ús i invertir esforços en iniciatives com Common Voice que puguin servir perquè la llengua no quedi enrere.

6 Conclusions

Durant els darrers vint anys, s'han assolit grans èxits pel que fa a la normalització del català a les tecnologies. El català està present en els principals sistemes operatius i programes, hi ha empreses que inverteixen diners i esforços per oferir els seus productes tecnològics en català i hi ha multitud de continguts a Internet en aquesta llengua. No obstant això, amb les noves innovacions tecnològiques torna a haver-hi el risc de perdre la partida enfront d'un nombre reduït de llengües dominants, per la qual cosa cal seguir treballant.

A més de la feina feta fins ara i les tecnologies lingüístiques desenvolupades fins al moment per impulsar la presència del català a l'àmbit tecnològic, que han de seguir servint de base, és important que siguem conscients de l'evolució tecnològica i les noves necessitats (per exemple, dades de veu) per avançar en la direcció correcta. Només aprofitant al màxim les inversions mitjançant formats i llicències lliures i identificant projectes clau en els quals implicar-se serà possible garantir la bona salut del català al món tecnològic.

Referències bibliogràfiques

Canals-Marote, Raul, Esteve-Guillen, Anna, Garrido-Alenda, Alicia, Guardiola-Savall, Maribel, Iturraspe-Bellver, Amaia, Montserrat-Buendia, Sandra, Ortiz-Rojas, Sergio, Pastor-Pina, Helder, Perez-Antón, Pedro M., i Forcada, Mikel L. (2001). The Spanish–Catalan machine translation system interNOSTRUM. Dins *Proceedings of MT Summit VIII, Santiago de Compostela* (p. 73-76).

Espanya. Real Decreto 2707/1985, de 27 de diciembre, por el que se declaran de obligado cumplimiento las especificaciones técnicas de los equipos teleimpresores, impresoras y máquinas de escribir electrónicas y su homologación por el Ministerio de Industria y Energía (BOE núm. 64 § 7017).

Forcada, Mikel L., Ginestí-Rosell, Mireia, Nordfalk, Jacob, O'Regan, Jim, Ortiz-Rojas, Sergio, Pérez-Ortiz, Juan Antonio, Tyers, i Francis M. (2011). Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, *25*(2), 127-144.

Ivars-Ribes, Xavier, i Sánchez-Cartagena, Víctor M. (2011). A Widely Used Machine Translation Service and its Migration to a Free/Open-Source Solution: the Case of Softcatalà. Dins *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*. Barcelona: Universitat Oberta de Catalunya.

Yahoo!, en català (10 de desembre de 2001). *324.* (Consultat el 30 de desembre de 2019).