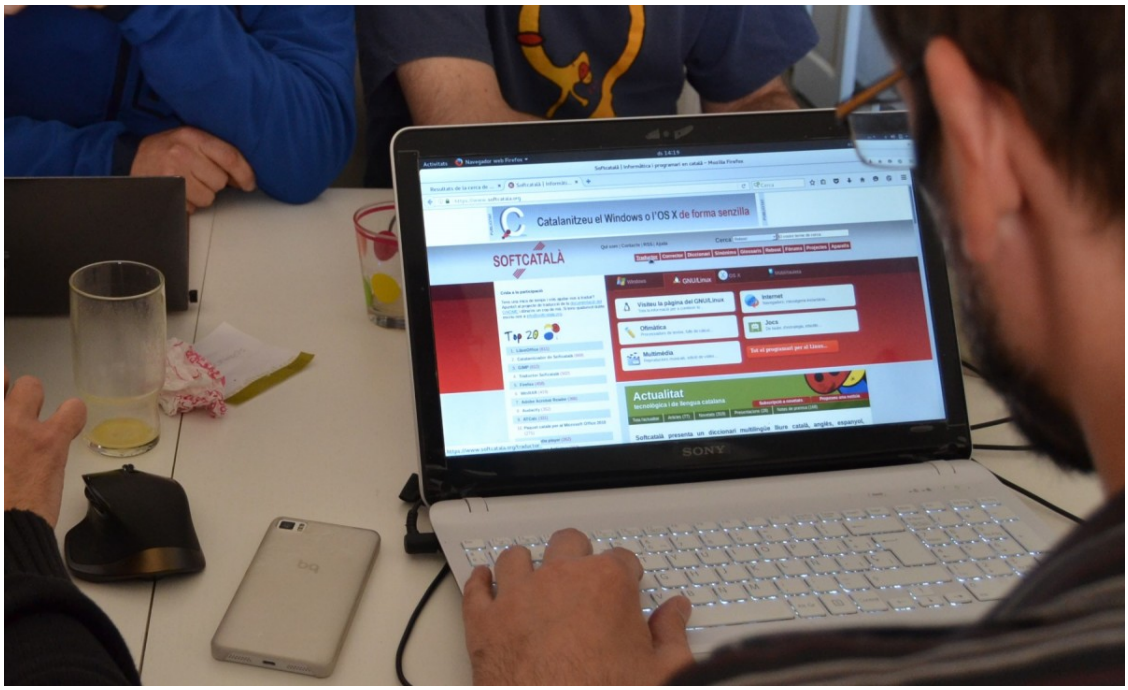
LLENGUA I LITERATURA | Baybars Külebi | Actualitzat el 29/06/2021 a les 09:20

El fet diferencial del català: la comunitat de programari lliure i obert

El punt més feble del català és el desenvolupament dels productes comercials d'escala gran | Una manera d'assegurar-ne la continuïtat és tenir una disposició de sobirania tecnològica, així com invertir en una infraestructura digital de l'administració pública



Entitats sense ànim de lucre com Softcatalà s'han distingit per impulsar la llengua en tots els àmbits digitals. | Anna B./Flickr

Sisè article del dossier «El català al món digital» (https://www.pensem.cat/grans/temes/1474/dossier-catala-al-mon-digital)

(https://www.pensem.cat/grans/temes/1474/dossier-catala-al-mon-digital)

Fa molts anys que diverses entitats impulsen la inclusió del català als productes del mercat. Els esforços (https://www.plataforma-llengua.cat/que-fem/noticies/4701/denunciem-que-les-versions-doblades-o-subtitulades-en-catala-amb-diners-publics-no-arriben-a-netflix) de la Plataforma per la Llengua no només són inavaluables sinó també ens dóna una memòria d'aquesta lluita. Una de les batalles d'aquesta lluita és la inclusió del català en els productes digitals. La raó d'acceleració de les activitats en aquesta línia està relacionat amb les innovacions importants en les tecnologies lingüístiques dels últims cinc anys. Amb la proliferació de l'ús de les xarxes neuronals els productes d'aprenentatge automàtic al mercat van augmentar. Des de traducció automàtica fins a la síntesi de la parla, aquestes eines faciliten les nostres vides i algunes són gairebé imprescindibles. Però sabem que la presència del català no és suficient encara. Un exemple molt important és, el fet que **cap dels 32 assistents de veu del mercat parla català**, segons l'InformeCat de 2020 (https://www.plataforma-llengua.cat/que-fem/estudis-i-publicacions/267/informecat-2020) .

Amb aquest article vull destacar els esdeveniments realitzats per abordar aquesta situació, a més fer-ho enfocant les activitats de la comunitat, sobretot la comunitat de programari obert i lliure. Perquè en alguns aspectes **les solucions proporcionades per la comunitat de programari obert i lliure estan per davant de les empreses privades** i també de les universitats per la seva aplicabilitat i accessibilitat. En la meva opinió, això és un fet diferencial que fa la llengua catalana especial, a més obre un **camí cap a una sobirania tecnològica** per assegurar la continuïtat de la presència de la llengua a l'àmbit digital. Abans de profunditzar els conceptes, comencem amb els fets i les novetats del camp.

[noticiadiariambautor]93/231[/noticiadiariambautor]
Quins són els punts forts i febles del català en l'àmbit digital?
El punt més fort del català és el seu potencial de mobilització de la comunitat. L'exemple més obvi són les activitats de Softcatalà i la seva mobilització de la comunitat per localitzar software obert i lliure. Gràcies a aquests esforços tenim eines imprescindibles (https://www.softcatala.org/projectes/) com Firefox, Libre Office i Ubuntu en català. A més, el projecte més impactant és l'impuls que van donar al Common Voice (https://commonvoice.mozilla.org/ca) , el projecte de Mozilla per recollir conjunts de dades de la parla, obert i lliure. Gràcies a la promoció de Softcatalà ara **el català representa una de les llengües més grans del corpus del Common Voice**. El català és la quarta llengua (després de l'anglès, l'alemany i el kinyarwanda) a la plataforma amb 755 hores d'enregistraments validades (dades de maig 2021).

A Col·lectivaT, la cooperativa sense ànim de lucre de la qual en sóc un dels cofundadors, tenim una altra estratègia, de generar conjunts de dades i eines obertes aprofitant els recursos ja existents. Durant la nostra curta vida vam crear dos conjunts de dades importants (https://collectivat.cat/rap) , el de TV3 i de ParlamentParla, aprofitant respectivament la

programació de TV3 i dels enregistraments del Parlament de Catalunya.

Actualment aquestes activitats estan alimentant tecnologies i prototips més concrets, l'exemple més destacat és assistent.cat (http://assistent.cat/) , que és el primer assistent virtual en català. És la versió localitzada de Mycroft (https://mycroft.ai/) , l'assistent virtual obert. Per ara Mycroft no proporciona un dispositiu en català (de fet no donen suport a cap altra llengua excepte anglès) i el desenvolupament de l'eina va ser possible a múltiples actors de la comunitat, des de membres de Softcatalà i Col·lectivaT a traductores voluntàries de Traumàtica de la UAB i d'altres desenvolupadors (https://www.softcatala.org/projectes/mycroft/) . A més, el desenvolupament d'aquest prova de concepte va ser possible gràcies a les altres tecnologies obertes clau, específicament el nostre Catotron (http://catotron.collectivat.cat/) , l'únic sistema de la síntesi de la parla en català basada en xarxes neuronals, i el sistema de reconeixement de la parla Vosk-Kaldi (https://alphacephei.com/vosk/models) , que té una precisió comparable, si no millor, que el servei de Google. Totes aquestes eines estan desenvolupades per la comunitat de programari lliure.

Cal destacar el suport del Departament de la Cultura durant els últims tres anys al desenvolupament d'aquestes eines i sobretot l'ampliació del potencial de la comunitat en si. Mitjançant el seu suport "per fomentar l'ús del català" estan donant un impuls imprescindible a la comunitat de programari lliure.

Dit tot això, des d'aquest punt de vista, **el punt més feble del català és el desenvolupament dels productes comercials d'escala gran**. Tot i que existeixen exemples dels productes destinats a l'usuari final com el traductor del Softcatalà (https://www.softcatala.org/traductor/) , fins ara no hi ha cap exemple de la difusió i adopció d'aquestes tecnologies recentment desenvolupades al mercat comercial en un producte concret. Aquest problema té múltiples facetes, i per entendre la manca dels productes innovadors al mercat abans de tot cal considerar la lògica del capital.

(https://www.pensem.cat/registre)

Com encaixa la promoció del català en el context d'un mercat global dominat per llengües majoritàries?

Aquí entrem en el tema de les empreses del mercat global i la seva falta d'interès en proporcionar

productes en català. La raó principal d'això és **la visió de les empreses per veure l'Estat espanyol com un sol mercat**. És a dir, per elles el suport de la llengua "més comuna" seria suficient per penetrar al territori peninsular. Això és la lògica del capital i de les empreses però no és una realitat ineludible.

Primer de tot, aquesta lògica està basada en **la sensació de les empreses que hi ha una falta de demanda dels productes en català**. Tot i que hi ha proves suficients de l'interès del mercat (https://llengua.gencat.cat/ca/detalls/noticia/Gairebe-un-90-dels-consumidors-volen-el-catala-en-els-assistents-de-veu) , a vegades no és tan evident fins que aparegui un projecte o esdeveniment amb un suport massiu. Un exemple recent és el cas del maori, la llengua indígena de Nova Zelanda. Després d'un projecte de la comunitat per recollir enregistraments dels parlants (https://www.wired.co.uk/article/maori-language-tech) , les empreses grans van començar a tenir molt d'interès en la llengua. Quan no podien comprar els drets de l'ús del conjunt de dades, van llançar un esforç per recopilar un conjunt de dades comercial. Actualment el propietari del conjunt de dades comunitari té finançament per desenvolupar una aplicació mòbil per facilitar l'aprenentatge de la llengua, amb una versió ja disponible.

Pel cas del català, sabem que la mateixa proveïdora -que probablement treballa per a Google- que gestiona la recopilació de dades de la parla va executar també un altre projecte per al català. Això ens mostra que **el català no està totalment oblidat per les grans empreses multinacionals**, per tant és una qüestió de temps fins que arribin els productes al mercat. Però hem de considerar bé si és la solució desitjable de portar les tecnologies avançades al mercat pel consum massiu.

[noticiadiariambautor]93/227[/noticiadiariambautor]
Com encaixa internament la promoció digital del català en un context de diglòssia a favor del castellà?
Després de parlar de les empreses grans multinacionals (o GAFAM; Google, Amazon, Facebook, Apple, Microsoft) i el seu possible interès en el català, cal parlar d'un altre tema pertinent a la situació del català dins l'Estat espanyol: **la sobirania tecnològica**.

Quan parlem de 'sobirania tecnològica' ens estem referint generalment a **control sobre les dades personals, control sobre els processos o algorismes que corren darrere dels serveis tecnològics i la possibilitat de reparar i/o modificar els dispositius**. Els productes de GAFAM infringeixen almenys un d'aquests principis; des de la impossibilitat de modificació dels dispositius d'Apple, a explotació de dades personals de Google fins als algorismes de Facebook que categoritzen les persones usuàries. A més d'aquests problemes genèrics, tenim un altre problema específic per a la situació del català com a llengua minoritzada: **la decisió d'oferir català o no en els serveis de les empreses grans és una prerrogativa seva**. La dependència a la voluntat d'aquestes empreses és el problema que el català està patint actualment. A més, fins i tot si decideixen integrar-lo, no hi ha cap garantia que aquests serveis es mantinguin en el futur. És a dir, la sobirania tecnològica no només implica control sobre l'ús dels productes tecnològics, sinó que també assegura la longevitat de les tecnologies desenvolupades. L'existència del català dins l'Estat espanyol sempre implicarà un cert perill per a la seva supervivència digital.

En el context dels productes tecnològics, és evident que la dominància del castellà sobre el català és per la lògica del capital i no per un estat opressiu ni per una societat discriminatòria, tot i que una alimenta l'altra. Però una de les maneres d'assegurar la proliferació contínua del català en l'àmbit digital és tenir una disposició de sobirania tecnològica. És innegable que el català és una llengua viva, amb una presència considerable als diversos mitjans, dels llibres a la producció audiovisual, i per això té tot el dret de tenir lloc en l'àmbit digital.

[noticiadiariambautor]93/224[/noticiadiariambautor]

Quines accions s'haurien de promoure perquè fos una llengua disponible a tots els serveis digitals?
Des d'aquest punt de vista de sobirania tecnològica, és **important apostar per tecnologies obertes i lliures**, impulsar la creació dels conjunts de dades obertes per a l'ús dels

desenvolupadors. L'aspecte més important de l'execució d'aquestes accions és el suport de la comunitat. La creació de les xarxes formals i informals de desenvolupadors, l'interès de les universitats per contribuir als projectes existents i l'organització de diverses activitats com *hackatons* són algunes de les accions més concretes per assegurar el desenvolupament continu de les tecnologies lingüístiques en català.

En aquest escenari hi ha un altre actor molt important, que és l'administració pública. Una altra manera d'assegurar el manteniment de les tecnologies desenvolupades i generar dades lingüístiques per millorar els productes tecnològics és **invertir a una infraestructura digital de l'administració pública**. L'ús de les tecnologies obertes en els serveis públics asseguraria el manteniment d'aquestes tecnologies. A més, cap a l'altra direcció, els serveis públics podrien ser fonts importants de dades per a la millora de les tecnologies lingüístiques, com el corpus de la parla del parlament ParlamentParla, que ja està essent utilitzat per múltiples projectes oberts. De fet ja hi ha un pla proposat per la Generalitat per invertir en una infraestructura de tecnologies lingüístiques (https://politiquesdigitals.gencat.cat/ca/detalls/Noticia/Nova-Noticia-00208) que preveu també una plataforma de dades lingüístiques.

[noticiadiariambautor]93/222[/noticiadiariambautor]
Quins actors o recursos caldria activar perquè fos possible?
Dins d'aquesta cadena de valor proposat, estem parlant d'alguns actors molt clars com les entitats de la comunitat i l'administració pública. Però queda una peça molt important que és **la comercialització de les tecnologies obertes desenvolupades**, que implica un altre tipus d'actor.

La comercialització fa referència al desenvolupament dels productes, des de l'escalabilitat de la tecnologia fins a la distribució i la capacitat d'atendre als clients. Aquestes tasques es poden dur a terme només per part de les entitats privades, i les que hi hauran d'apostar són les PIMES del territori, és a dir, els fabricants que tenen un arrelament territorial.

D'aquesta manera, podrem **desenvolupar una xarxa dels actors tecnològics, una xarxa sobirana i a més resistent**, en el sentit que el manteniment de les tecnologies estarà assegurat, per la implicació de les entitats privades locals.

Aquesta estratègia és important per no dependre de les grans empreses multinacionals. Fa anys que la Plataforma per la Llengua i l'administració pública estan pressionant les empreses de GAFAM per integrar el català, però els resultats són encara molt limitats. Apel·lar a l'autoritat no ha donat un resultat concret en els productes de la parla encara, però mentrestant la comunitat està creant noves tecnologies, prototips i productes senzills. **Ara cal activar iniciatives comercials al territori i fer-les escalables**. Això no només omplirà un buit en la provisió dels serveis (l'oferta), sinó que també impulsarà les empreses grans indirectament perquè es mostrarà definitivament la preferència dels catalanoparlants pels serveis en la seva llengua, és a dir una demanda concreta.

En resum, a més de la inversió, **cal assumir una estratègia de sobirania tecnològica** que implica els actors des de la comunitat del programari lliure i les iniciatives privades, fins a l'administració pública. La innovació en obert té el potencial d'impulsar iniciatives locals, motivar empreses grans i també dinamitzar la comunitat per adoptar i mantenir aquestes tecnologies clau. Durant els últims 3 anys el català ha fet un salt bastant important, però **ens queda un pas més per portar l'experiència tecnològica i els productes nous al mercat**.

[noticiadiariambautor]93/223[/noticiadiariambautor]
Referències d'interès:

Vosk és un projecte de programari lliure per fer el motor del reconeixement de la parla Kaldi més accessible. Està impulsat per Alphacephei i el contribuidor principal és Nikolay V. Shmyrev. Els models en català estan a la seva pàgina https://alphacephei.com/vosk/models (https://alphacephei.com/vosk/models)

Els models del motor de reconeixement de la parla entrenats per Ciaran O'Reilly estàn disponibles a https://huggingface.co/ccoreilly/wav2vec2-large-xlsr-catala (https://huggingface.co/ccoreilly/wav2vec2-large-xlsr-catala) i https://github.com/ccoreilly/wav2vec2-catala (https://github.com/ccoreilly/wav2vec2-catala)

Articles del dossier:
[noticiadiariambautor]93/234[/noticiadiariambautor]
[noticiadiariambautor]93/233[/noticiadiariambautor]
[noticiadiariambautor]93/229[/noticiadiariambautor]
[noticiadiariambautor]93/232[/noticiadiariambautor]
[noticiadiariambautor]93/226[/noticiadiariambautor]
[noticiadiariambautor]93/231[/noticiadiariambautor]
[noticiadiariambautor]93/230[/noticiadiariambautor]
[noticiadiariambautor]93/227[/noticiadiariambautor]
[noticiadiariambautor]93/223[/noticiadiariambautor]
[noticiadiariambautor]93/224[/noticiadiariambautor]
[noticiadiariambautor]93/222[/noticiadiariambautor]
[noticiadiariambautor]93/225[/noticiadiariambautor]