

## Grans tecnològiques, demanda i igualtat lingüística, claus per al català

*La situació actual de la llengua és força desfavorable en un àmbits essencial com són les tecnologies del llenguatge | La relació amb les grans empreses tecnològiques sempre és un punt de controvèrsia, però cal estar-hi amatent*



La seu central de Microsoft, a Redmond, Washington (EUA). | Microsoft.

*Vuitè article del dossier «El català al món digital»*



Quins són els punts forts i febles del català en l'àmbit digital?

Històricament, el català ha tingut una llarga trajectòria de superació de reptes complexos, com per exemple es va viure durant l'aparició d'Internet, el desplegament de la web (i la web 2.0), els dispositius mòbils o les xarxes socials, on el català s'ha pogut anar consolidant durant tot aquest temps i tenir una certa rellevància. Actualment, quan s'està produint el **desplegament de l'anomenada quarta revolució industrial** impulsada per la Intel·ligència Artificial (IA) i les tecnologies del llenguatge, moltes llengües -minoritàries o no- s'enfronten a un nou repte.

Per identificar el que considerem fortaleeses del català en el món digital cal destacar l'impuls que des de l'any 2000 en endavant s'ha donat a tot l'àmbit d'Internet i el món web en general. **El català té actualment una presència notable a Internet, tot i ser una llengua mitjana en l'àmbit europeu.** Aquest impuls passat ha de ser una referència per a la nova revolució tecnològica que està actualment en desplegament. A més a més, tots aquests continguts generats en català ens permeten disposar d'una àmplia base de textos, documents, àudios i imatges amb les quals treballar per generar corpus per l'entrenament i millorar algorismes i motors de llenguatge natural, reconeixement automàtic de la veu, generació de veu, traducció automàtica, etc.

En la nova fase de la transformació digital que s'ha iniciat, la IA serà el motor que la impulsarà i on les tecnologies del llenguatge hi tindran una importància cabdal. Sorgeix aquí el **punt de major debilitat del català**, com també el d'altres llengües molt més parlades, a causa de l'ús dominant de la llengua anglesa.

Analitzem com a cas il·lustratiu el motor de **GPT-3**, el revolucionari model de llenguatge presentat a principis de 2020 i que fa ús de l'aprenentatge profund per produir textos que simulen reaccions humanes. Si estudiem el document que detalla el model GPT-3 veiem que va ser pre-entrenat amb un gran conjunt de dades que inclouen el popular conjunt de dades Common Crawl, format per un **gran volum de petabytes de dades** recollides des de 2008. Aquest *data set* representa al voltant del 60% de la mescla total de l'entrenament de GPT-3, que també inclou d'altres conjunts de dades com WebText2, Books1, Books2, etc.

Segons la informació de GitHub de Common Crawl, aquesta és la distribució dels idiomes en els documents del conjunt de dades:

## Statistics of Common Crawl Monthly Archives

Number of pages, distribution of top-level domains, crawl overlaps, etc. - basic metrics about Common Crawl Monthly Crawl Archives  
Latest crawl: CC-MAIN-2021-21

[Home](#)

[Size of crawls](#)

[Top-level domains](#)

[Registered domains](#)

[Crawler metrics](#)

[Crawl overlaps](#)

[Media types](#)

[Character sets](#)

[Languages](#)

[View the Project on GitHub](#)

This project is maintained by [commoncrawl](#)

Hosted on GitHub Pages — Theme by [orderedlist](#)

## Distribution of Languages

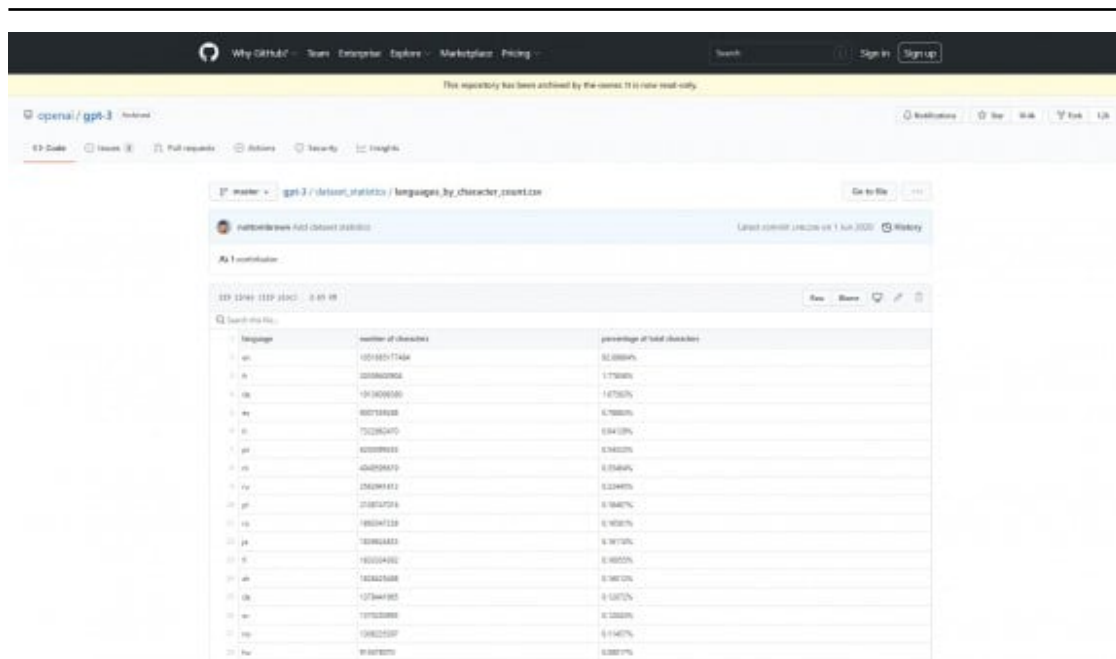
The language of a document is identified by Compact Language Detector 2 (CLD2). It is able to identify 160 different languages and up to 3 languages per document. The table lists the percentage covered by the primary language of a document (returned first by CLD2). So far, only HTML pages are passed to the language detector. The underlying data including page counts is provided in [languages.csv](#).

crawl	CC-MAIN-2021-10	CC-MAIN-2021-17	CC-MAIN-2021-21
language	%	%	%
eng	44.2918	43.9003	44.8424
rus	7.4865	7.6919	7.2658
zho	5.8281	5.3239	4.6508
deu	5.4853	5.7249	5.6906
jpn	4.6689	4.9134	4.5850
fra	4.4390	4.5278	4.4747
spa	4.3295	4.3587	4.3315
ita	2.3657	2.4426	2.4277
<unknown>	2.2552	2.0984	2.8840
por	2.1705	2.1740	2.1373
nld	1.7745	1.8616	1.8082
pol	1.5764	1.6096	1.6050
ces	1.1067	1.1496	1.1124
tur	1.0209	1.0361	1.0099
vie	0.8387	0.8630	0.8036
ind	0.8278	0.8610	0.8311
swe	0.7256	0.7461	0.7398

*L'anglès és hegemònic a l'eina essencial utilitzada per pre-entrenar el nou sistema d'intel·ligència artificial GPT-3, Common Crawl.*

L'anglès és clarament la llengua dominant en el corpus, ja que representa més del 44% del total de documents, mentre que la segona llengua més representada és el rus, a gran distància amb menys d'un 7% del total. Algunes de les 10 llengües més parlades en termes de parlants nadius estan clarament infrarepresentades. L'espanyol, per exemple, segona llengua amb 460 milions de parlants nadius (més que els parlants nadius d'anglès), només representa el 4% del total de documents. **La posició del català també és molt residual, només representa 0,22 % de les entrades (posició 32a).**

A la següent estadística s'hi pot visualitzar el nombre total de caràcters per idioma en el conjunt de dades d'entrenament GPT-3:



The screenshot shows a GitHub repository page for 'openai/gpt-3'. A yellow banner at the top states 'This repository has been archived by the owner. It is now read-only.' Below the repository name, there are navigation links for Code, Issues, Pull requests, Actions, Security, and Insights. The main content area shows a file named 'languages\_by\_character\_count.csv' with a commit history of 1. The file content is a table with three columns: 'language', 'number of characters', and 'percentage of total characters'. The table lists 20 languages, with 'en' having the highest count and percentage.

language	number of characters	percentage of total characters
en	10148217434	52.0884%
fr	2288622824	11.7528%
de	1914029340	9.7503%
es	882134828	4.5088%
it	722262472	3.6413%
pt	622222222	3.1422%
ru	422222222	2.1422%
sv	282222222	1.4222%
pl	222222222	1.1222%
uk	182222222	0.9222%
ja	122222222	0.6222%
nl	102222222	0.5222%
sk	822222222	0.4222%
uk	722222222	0.3622%
ar	622222222	0.3222%
he	522222222	0.2622%
hu	422222222	0.2222%

Nombre total de caràcters per idioma en el conjunt de dades d'entrenament GPT-3.

Si ens fixem en una altra eina molt utilitzada en solucions de llenguatge natural, *Google Bidirectional Encoder Representations and Transformers* (més conegut com a BERT), en la seva plana de GitHub informen que dona suport als 100 idiomes principals que hi ha representants en els sets de dades de Wikipedia. Fent una avaluació detallada del sistema, només s'ha fet un refinament real en els 15 idiomes. Tot i que BERT tècnicament admet més idiomes, el menor nivell de precisió i la manca de proves adequades limita l'aplicabilitat d'aquesta tecnologia.

En resum, mentre que la gran quantitat de dades d'entrenament fa que els models lingüístics com el GPT-3 o BERT siguin bons en diversos idiomes, tot i estar optimitzats per a l'anglès i/o utilitzen un petit percentatge de documents no anglesos, **les llengües minoritàries (fins i tot moltes de les no minoritàries) continuen estant en gran desavantatge.**

Atès l'èxit en el desenvolupament gràcies a les entitats públiques, privades i de programari lliure per la promoció i posicionament del català a Internet, i tot i que la situació actual de les tecnologies del llenguatge envers el català no és la desitjada, creiem que **disposem d'una base prou sòlida sobre la qual treballar un estratègia d'innovació, investigació i d'inversió** que hauria de portar-nos a disposar d'eines, models, i finalment de solucions, en les quals el català tingui les mateixes oportunitats que altres idiomes més parlats.

Com a conclusió final podem pensar que la **situació actual del català és força desfavorable** en un dels àmbits principals pel desplegament de la IA, sobretot a causa de la situació dominant de l'anglès. Tot i això, hem vist que en un passat no molt llunyà -on també van aparèixer reptes tecnològics importants que podien semblar també complexos d'assolir- diferents grups, associacions, entitats públiques i també privades van obtenir recursos i dur a terme accions per fer del català una llengua viva, d'utilitat i amb futur. Això precisament és el que ens cal agafar com a referència i projectar cap al futur: nous models per treballar de forma adequada i efectiva dins de l'àmbit de les tecnologies del llenguatge, que ja hem identificat com el motor que impulsarà el nou paradigma guiat per la IA, que voldrem que parli i entengui la nostra llengua.

Quines accions s'haurien de promoure perquè fos una llengua disponible a tots els serveis

# Pensem.

---

digitals?

En la fase de transformació digital actual, els serveis digitals aniran millorant proporcionant una experiència d'usuari cada cop més simple, més propera i en general més natural. Això ha de permetre la reducció de les barreres 'digitals' que fins ara alguns d'aquests serveis podien aixecar, com són la configuració de múltiples dispositius, la complexitat cada cop més gran en l'ús del mòbil, etc.

Part d'aquesta simplificació serà causada per la forma d'accedir, consumir i d'interactuar amb els serveis digitals, la qual farà ús de forma intensiva de serveis basats en tecnologies del llenguatge com l'ASR (Reconeixement automàtic de la parla), TTS (text a veu), NLP (Processament de llenguatge natural) i NLG (generació de llenguatge natural). Aquestes tecnologies es combinaran per generar l'aspecte d'un humà digital. L'empresa Soul Machines, per exemple, ja ofereix serveis de disseny i creació de persones digitals a mida de les necessitats dels clients.



*Una internauta interacciona amb una assistent digital. Foto: Soul Machines*

**La disponibilitat de l'ús del català en aquestes tecnologies requerirà d'un pla d'accions** que caldrà definir, planificar, i executar en diferents àmbits i actors, si és possible en paral·lel, buscant sinergies i col·laboracions creuades. Dins d'aquestes accions hi identifiquem de forma prioritària:

## **1) Foment de la relació amb grans empreses tecnològiques, sobretot amb les més implantades a Catalunya**

La relació amb grans empreses tecnològiques sempre és un punt de molta controvèrsia. Tot i que esperar que la seva estratègia permeti disposar de serveis per a relacionar-nos en català no hauria de ser l'opció principal, sí que cal seguir atents als seus moviments, i no deixar de reclamar, tant des de les institucions públiques com privades, un tracte per a la llengua catalana - recordem que és parlada i compresa per més de 10 milions d'habitats-, el mateix tracte que tenen d'altres llengües també minoritàries. Per posar alguns exemples: **Google ha implementat idiosincràcies regionals específiques en diferents regions d'Europa** (com a Suïssa per l'alemany o el francès), o com Apple ha incorporat a Siri els idiomes finès, hebreu, danès o noruec, tots ells amb menys parlants que el català. Després hi ha els casos dels diferents

altaveus intel·ligents, que actualment no tenen planificat el futur suport del català.

Tot i així, aquestes grans companyies sí que proporcionen d'altres serveis de més baix nivell que ens poden permetre construir un marc de solucions tecnològiques per a la implementació de solucions d'assistents en català. A part del conegut i molt utilitzat Google Translate, d'altres companyies com **Microsoft** ja ofereixen serveis avançat de qualitat en català, com per exemple la solució d'ASR i TTS a la seva plataforma *cloud* Azure. Aquests serveis estan disponibles tant pel servei de STT amb detecció automàtica de català com el serveis de TTS on hi ha **disponibles fins a tres veus neurals en català**.

such as e-books into audiobooks and enhance in-car navigation systems. With the human-like natural prosody and clear articulation of words, neural voices significantly reduce listening fatigue when users interact with AI systems.

Note  
Neural voices are created from samples that use a 24 kHz sample rate. All voices can upsample or downsample to other sample rates when synthesizing.

Language	Locale	Gender	Voice name	Style support
Arabic (Egypt)	ar-EG	Female	ar-EG-SalwaNeural	General
Arabic (Egypt)	ar-EG	Male	ar-EG-SherkhanNeural	General
Arabic (Saudi Arabia)	ar-SA	Female	ar-SA-ZahraNeural	General
Arabic (Saudi Arabia)	ar-SA	Male	ar-SA-HamadNeural	General
Bulgarian (Bulgaria)	bg-BG	Female	bg-BG-RitaNeural	General
Bulgarian (Bulgaria)	bg-BG	Male	bg-BG-KirilNeural	General
Catalan (Spain)	ca-ES	Female	ca-ES-ElviraNeural	General
Catalan (Spain)	ca-ES	Female	ca-ES-SonataNeural	Catalan
Catalan (Spain)	ca-ES	Male	ca-ES-PereCristianNeural	Catalan
Chinese (Cantonese, Traditional)	zh-HK	Female	zh-HK-LisaNeural	General

Microsoft contempla tres veus neurals en català.

## 2) Potenciar el desenvolupament de les tecnologies del llenguatge

**La disponibilitat de serveis de llenguatge natural en un idioma com el català pot arribar a ser crític per al seu futur**, amb el risc de ser substituïts cada vegada més en els serveis en línia pels quals s'utilitzen de forma més general i, per tant, optimitzats en les noves solucions d'aprenentatge automàtic.

Per això, acadèmies de la llengua com la RAE, el govern d'Espanya o d'altres institucions com la Generalitat de Catalunya o el govern basc estan llançant iniciatives específiques per a abordar aquest problema:

Plan de Impulso de las Tecnologías del Lenguaje. Aquest Pla té per objectiu fomentar l'impuls de les tecnologies del llenguatge, traducció automàtica i sistemes conversacionals tant per la llengua espanyola com totes les cooficials.

La RAE ha presentat el projecte LEIA per a "*aprovechar la inteligencia artificial para crear herramientas que fomenten el uso correcto del español en los seres humanos*". En el projecte hi ha la participació de Telefónica així com empreses tecnològiques com Google, Amazon, Microsoft, Twitter i Facebook.

Al País Basc, el Basque Center for Language Technology és l'entitat dins de l'àmbit de les

institucions públiques i de les universitats que treballa en proporcionar solucions i actius per poder disposar de la llengua eusquera en l'àmbit digital.

A nivell de Catalunya, són diverses les iniciatives que s'han creat durant l'últim any i mig per impulsar les tecnologies del llenguatge:

**Catalonia.ai.** És un marc que agrupa les actuacions de suport a l'evolució de la IA a Catalunya. Està coordinat pel Departament de Polítiques Administració Pública del Govern de la Generalitat de Catalunya. Té l'objectiu de construir un ecosistema d'intel·ligència artificial on cooperaran la comunitat científica i acadèmica, les empreses del sector, i els ciutadans per arribar a desenvolupar una indústria rellevant que situï Catalunya i Barcelona entre els motors d'IA del continent i del món.

**Projecte Aina.** És un projecte del departament de Polítiques Digitals per la creació de recursos digitals i lingüístics necessaris per facilitar el desenvolupament d'aplicacions basades en la intel·ligència artificial i les tecnologies de la llengua, com ara els assistents de veu, els traductors automàtics o els agents conversacionals en català.

**CIDAI.** Segueix el model de *Digital Innovation Hubs* definit per la Comissió Europea i es configura com a un centre en xarxa al servei de les empreses i també Instituciones. Es una peça important dins de la pròpia estratègia definida en el marc de Catalonia.ai.

### 3) Generació de demanda de productes i serveis digitals en català

Totes les accions identificades fins aquí no agafaran el nivell d'inèrcia necessari per iniciar un cercle virtuós -que realmenti tota aquesta la inversió en productes digitals en català i faci el model sostenible- si no **es genera una demanada al mercat**. En l'entorn multilingüista on el predomini de l'anglès és tan important, aquesta demanada haurà de ser impulsada per accions de diferent índole:

Incrementar l'oferta i potenciar l'augment de productes i serveis digitals en català.

Afavorir el consum i l'ús de productes i serveis digitals en català.

Promocionar els productes i serveis digitals que incorporen la llengua catalana amb models d'èxit.

Impulsar el català en les interfícies de veu i la intel·ligència artificial de mercat així com construir i desplegar un ventall d'assistents desenvolupats per les mateixes organitzacions tan públiques com privades.

Com ja hem comentat, impulsar de forma urgent el desenvolupament de les tecnologies lingüístiques basades en el català.

Com a conclusió final, volem contraposar l'impacte sociocultural en el cas de no treballar en les accions que segons el nostre criteri caldria per promocionar el català. És clar veure la tendència i l'impacte que els nous serveis digitals, la seva difusió, la simplificació en el seu ús que vindrà de la mà de les tecnologies del llenguatge que hem anomenat.

**«La situació d'inferioritat actual del català es pot veure accentuada per l'entorn multilingüe, impactant en les diferents generacions que identificaran el català com una llengua residual, no útil per interactuar en el món digital»**

La situació d'inferioritat actual del català es pot veure accentuada per l'entorn multilingüe en el qual convivim, impactant en les diferents generacions que ja són nadius digitals i que identificaran el català com una **llengua residual**, no útil per interactuar en el món digital.

Una llengua avança i es desenvolupa a mesura que es fa servir i és important tenir en compte que la tecnologia només sigui accessible quan les seves eines estan disponibles en el seu idioma. En un nivell bàsic, la falta de tecnologia de correcció ortogràfica afecta a qui parla i escriu idiomes menys comuns. Aquesta disparitat augmenta en la cadena tecnològica. Això significa que **a mesura que les tecnologies del llenguatge continuïn desenvolupant-se sense incorporar el català, serà més difícil incorporar-los en el futur**, posant en perill la seva evolució.

[noticiadiariambautor]93/222[/noticiadiariambautor]

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

La globalització i el gran domini que tenen en l'àmbit de les tecnologies de la informació les llengües més parlades -i l'anglès com a gran llengua dominant- no ha fer-se servir com una justificació per a desatendre la gran herència lingüística, cultural i social que tenim en tota la regió de parla catalana, de més de 10 milions de parlants. Tot al contrari, hauria de ser el que ens motivi a engagar les accions i el plans necessaris per disposar de les eines que requerim. Durant gran part de la història de la llengua catalana **han sorgit reptes que han pogut ésser superats**, arribant a la situació actual.

**Estem vivint l'era digital**, una etapa de transformació on, en qualsevol moment i en qualsevol lloc, per mitjà de tot tipus de dispositius, tenim accés a serveis digitals: web, xarxes socials, vídeo, etc. La interacció amb aquests dispositius és cada vegada més transparent: mentre que les primeres aplicacions comercials eren comandades únicament per teclat, avui en dia, gràcies als avenços de les tecnologies de la llengua, podem usar la nostra pròpia veu de forma quasi natural.

Però **si volem poder fer ús del català requerirem de solucions basades en tecnologies de la llengua** (reconeixement de veu, síntesi de veu, traductors, etc.) i recursos lingüístics apropiats (corpus orals i textuais, etc.), el desenvolupament dels quals haurà d'anar lligat a importants recursos per tal de generar la tecnologia necessària.

Podem posar com a exemple una història potser no molt coneguda. **La Xina es va enfrontar a un enorme problema als anys 80**, quan es va adonar que els seus més de 70.000 caràcters no cabien en un teclat. Si no haguessin resolt aquest problema amb el 'mètode Wubi', seguiria sent la Xina una superpotència tecnològica 40 anys després? Què passarà d'aquí a 40 anys amb les llengües que no rebin suficient suport de les tecnologies del llenguatge?



## 86版五笔字根表

金夕夕夕夕夕 夕夕夕夕夕夕 夕夕夕夕夕夕 夕夕夕夕夕夕	人入イ厂 八ハ残狄	月多乃乃 目目册用 ハマヨ永 氏氏衣家	白勺匕 手尹才丰 斤斤厂	禾ノ一ノ 禾ノ竹竹 才女女	言讠讠讠 言讠讠讠 广文方古 圭章	立ミマ< 立守六> 幸广斗才 立ノ门門	水彡水水 余余火>< 小小小木 业业业立 少	火灠灠米 亦小业恭	之讠讠 灠灠才良
35 Q 我	34 W 人	33 E 有	32 R 的	31 T 和	41 Y 主	42 U 产	43 I 不	44 O 为	45 P 这
工戈弋七 厂左七七 匸匸匸匸 艹井甘廿 艹	木木丁丁 西面西	大三手手 犬ナナフ 厂古石長 巨	土二土干 寸寸十串 雨灠灠	王一ノ 圭五戈	目丨丨 且卜丨上 止止产产	日丨丨日 日丨丨日 白四早虫	口川川	田囙血口 囙囙囙四 甲甲車車 力车	
15 A 工	14 S 要	13 D 在	12 F 地	11 G 一	21 H 上	22 J 是	23 K 中	24 L 国	
35 ← 31 41 → 45 3区(前左毛) 4区(前左毛)	15 ← 11 21 → 24 1区(前左毛) 2区(前左毛)	55 ← 51 25 5区(前左毛) 键盘分区图	多玄玄赤 弓弓口片 匕匕匕匕 豨豨	又巴马馬 スマムム	女ㄩ刀九 ヨヨヨヨ 白白	子ㄩ子了 也耳卩卩 口口卩卩	巳乙巳巳 コヨ尸尸 心寸小尸 羽门才才 羽上上上	山由巛巛 冂门门冂 凡凡凡凡 凡凡	
			55 X 经	54 C 以	53 V 发	52 B 了	51 N 民	25 M 同	

蓝字根：键名字  
红字根：笔画标识字根  
蓝字根：用于繁体字  
下方蓝字：区位、按键  
下方蓝字：一级简码字

字 根 口 诀	11G 王旁青头戋五一	21H 目具上止卜虎皮	31T 禾竹一撇双人立， 反文条头共三一	41Y 言文方广在四一	51N 已半巳满不出己
	12F 土土二千十寸雨	22J 日早两竖与虫依	32R 白手看头三二斤	42U 立辛两点六门广	52B 子耳了也框向上
	13D 太三(羊)古石厂	23K 口与川，字根稀	33E 月多(衫)乃用家衣底	43I 水旁兴头小倒立	53V 女刀九白山朝西
	14S 木丁西	24L 田甲方框四车力	34W 人和八，三四里	44O 火业头，四点米	54C 又巴马，丢矢矣
15A 工戈草头右框七	25M 山由巛，下框几	35Q 金勺缺点无尾鱼， 犬旁留儿一点夕， 氏无七(妻)	45P 之宝盖，摘(示) 示(衣)	55X 慈母无心弓和匕， 幼无力	

Mètode Wubi per distribuir els 70.000 caràcters xinesos en un teclat estàndard. Foto: Wikipedia

Analitzarem ara l'encaix de la promoció del català basant-nos en els tres àmbits geo-polítics dins dels quals està inclòs:

**Àmbit dels territoris de parla catalana.** La generació de demanda i dinamització en la àrea de parla catalana (Catalunya, País Valencià i les Illes) considerants també les diferents variants lingüístiques. La promoció de l'ús de la llengua ha d'anar dirigit a:

- Increment de l'oferta i potenciar l'augment de productes i serveis digitals en català.
- Afavorir el consum i l'ús de productes i serveis digitals en català.
- Promocionar els productes i serveis digitals que incorporen la llengua catalana amb models d'èxit.
- Impulsar el català en les interfícies de veu i la intel·ligència artificial de mercat, així com construir i desplegar un ventall d'assistents desenvolupats per les mateixes organitzacions tant públiques com privades.
- Com ja hem comentat, impulsar de forma urgent el desenvolupament de les tecnologies lingüístiques basades en el català.

**Àmbit estatal.** Tant des de les institucions públiques del govern d'Espanya com de les empreses privades voldran proporcionar i disposar de solucions compatibles amb el català per tal d'oferir

serveis de qualitat més propers als ciutadans. Serà molt important tot allò que s'ha comentat anteriorment sobre la generació de demanda per part dels ciutadans i clients també en aquest àmbit. Dins del govern d'Espanya destaca el *Plan de Impulso de las Tecnologías del Lenguaje*. Aquest pla té per objectiu fomentar l'impuls de les tecnologies del llenguatge, traducció automàtica i sistemes conversacionals tant per la llengua espanyola como totes les cooficials.

**La dinamització de l'àmbit privat també ha de ser un objectiu prioritari.** Si posem com a exemple els quatre principals bancs de l'estat, tots ells disposen d'accés web en català i l'ofereixen altres serveis com els caixers automàtics. De forma natural en el moment que sigui possible tècnicament i econòmicament oferir serveis mitjançant l'ús del català, per text o veu, aquests estaran disponibles. D'altres grans empreses del sector de la moda o serveis de telecomunicació també ofereixen els serveis en múltiples idiomes en canals com la web o atenció telefònica, per tant també estan interessats en aquests serveis.

**Àmbit de la Unió Europea.** La UE forma un conjunt de ciutadans de més de 500 milions de persones que comparteixen unes 80 llengües diferents. El **multilingüisme** a Europa és un fet diferencial clau per enfortir la integració entre totes les regions. La Unió Europea va adonar-se de la importància de les tecnologies de la llengua com a motor de la unitat europea i ha anat finançant diferents projectes en aquest àmbit. Un dels més rellevants aprovats és *Language equality in the digital age*, que planteja la coordinació a gran escala per a la recerca, desenvolupament i innovació en l'àmbit de les tecnologies del llenguatge, així com assegurar un lideratge en AI (dominat principalment pels EUA i la Xina).

La idea de la **igualtat lingüística digital (ILE)** és proporcionar un escenari on totes les llengües tenen el suport tecnològic i el context necessaris per seguir existint i prosperar com a llengües vives en l'era digital. D'aquest projecte inicial neix el projecte ELE (European Language Equality). El projecte té l'objectiu de definir l'estratègia de recerca, innovació i aplicació per aconseguir la **plena igualtat lingüística digital a Europa d'aquí a 2030**.

Un dels primers objectius del projecte és generar uns indicadors o mètriques sobre la igualtat lingüística. Aquestes mesures hauran de reflectir la preparació digital d'una llengua i la seva contribució a l'estat del multilingüisme facilitat per la tecnologia. La mètrica DLE (PDF) es calcula per a cada llengua en funció de diversos factors, agrupats en suport tecnològic (factors tecnològics, per exemple, els recursos, eines i tecnologies lingüístiques disponibles) i una sèrie de factors de context de la situació (per exemple, socials, econòmics, educatius, industrials).

El projecte d'igualtat lingüística disposa de múltiples fases i varis terminis. En la fase actual s'estan generant les bases, definicions i mètriques per a després iniciar els estudis de les diferents llengües minoritàries per poder fer l'anàlisi posterior. La participació en l'informe del català estarà liderat pel Barcelona Supercomputing Center i l'equip que també participa en el projecte AINA amb el Departament de polítiques digitals de la Generalitat de Catalunya.

Es d'esperar que els 3 àmbits detallats es combinin generant sinergies que permetin accelerar la innovació tecnològica, la posada en marxa de serveis i un increment de la demanda per part dels ciutadans i clients de serveis digitals. En el cas que no s'aconsegueixi incloure les llengües minoritàries (i no tant minoritàries) pot impactar en què la IA i els sistemes que amb els que aquesta tecnologia es desenvolupen estaran basats principalment en l'anglès, amb el que tot el que això pot representar en:

Evolució i reforç de les llengües: una llengua avança i es desenvolupa a mesura que es fa servir, així com tenir en compte que la tecnologia només es accessible quan les seves eines estan disponibles en el teu idioma. En un nivell bàsic, per exemple, no disposar de tecnologia de correcció ortogràfica afecta només a qui parla i escriu els idiomes menys comuns. Aquesta disparitat augmenta en la cadena tecnològica. Això significa que a la mesura que les tecnologies del llenguatge (reconeixement de la parla, comprensió del llenguatge, etc.) continuïn desenvolupant-se sense incorporar el català, serà més difícil incorporar-los en el futur, posant en perill la seva evolució.

Biaix cultural i normatiu. L'anglès i els idiomes adjacents no són representatius d'altres idiomes del món, ja que tenen estructures gramaticals úniques que moltes llengües no comparteixen. D'aquesta forma, el suport a Internet i altres tecnologies a l'anglès, progressivament el consideren com l'idioma predeterminat. Com que un sistema inicialment agnòstic es construirà de base en anglès, aprèn les normes i els sistemes d'una llengua específica, amb totes les implicacions culturals que comporten aquesta limitació. Aquest enfocament només continuarà fent-se més evident a mesura que les tecnologies del llenguatge no s'apliquin processos més intel·ligents que tinguin un enfocament internacional.

Evolució i millora dels algoritmes. Quan apliquem tècniques d'aprenentatge automàtic a només un grapat de llenguatges, estem programant biaixos implícits als sistemes. Com que l'aprenentatge automàtic i les tecnologies del llenguatge continuen avançant, encara que només en alguns idiomes, no només fem més difícil la introducció de nous idiomes, sinó que correm el risc de fer-ho impossible. Per exemple, la implementació de *tokenització* de paraules (procés molt utilitzat en processament de llenguatge natural que divideix cadenes llargues en peces més petites o *tokens*) té un rendiment molt baix en alguns idiomes, per exemple, els que presenten una reduplicació, una característica comuna a moltes llengües internacionals com l'afrikaans, l'irlandès, el panjabi i l'armeni.

**La importància d'incloure les llengües minoritàries està començant a ser una prioritat**, tal i com hem pogut comprovar, però no només per un tema cultural, sinó també per el seu impacte social i tecnològic.

#### Bibliografia rellevant:

Gaspari, F. et al (2021). "Digital Language Equality (preliminary definition)" (PDF). European Language Equality (ELE)  
Pereira, D. (2021). "How can NLP impact the future of minority languages?"  
Plataforma per la Llengua (2020). "InformeCAT 2020" (PDF)  
Siavoshi, M. (2020). "The Importance of Natural Language Processing for Non-English Languages"

#### 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]