LLENGUA I LITERATURA | Marta Villegas | Actualitzat el 10/06/2021 a les 09:05

El català i més de 20 llengües europees, enfront de l'extinció digital

«Fa un any hagués dit que la tecnologia jugava al nostre favor; avui, ja no ho tinc tan clar» | «La promoció de la llengua pròpia és una preocupació de totes les llengües no majoritàries, que són la immensa majoria»

El Parlament Europeu va aprovar el 2018 una resolució sobre la igualtat lingüística a l'era digital. | Parlament Europeu.

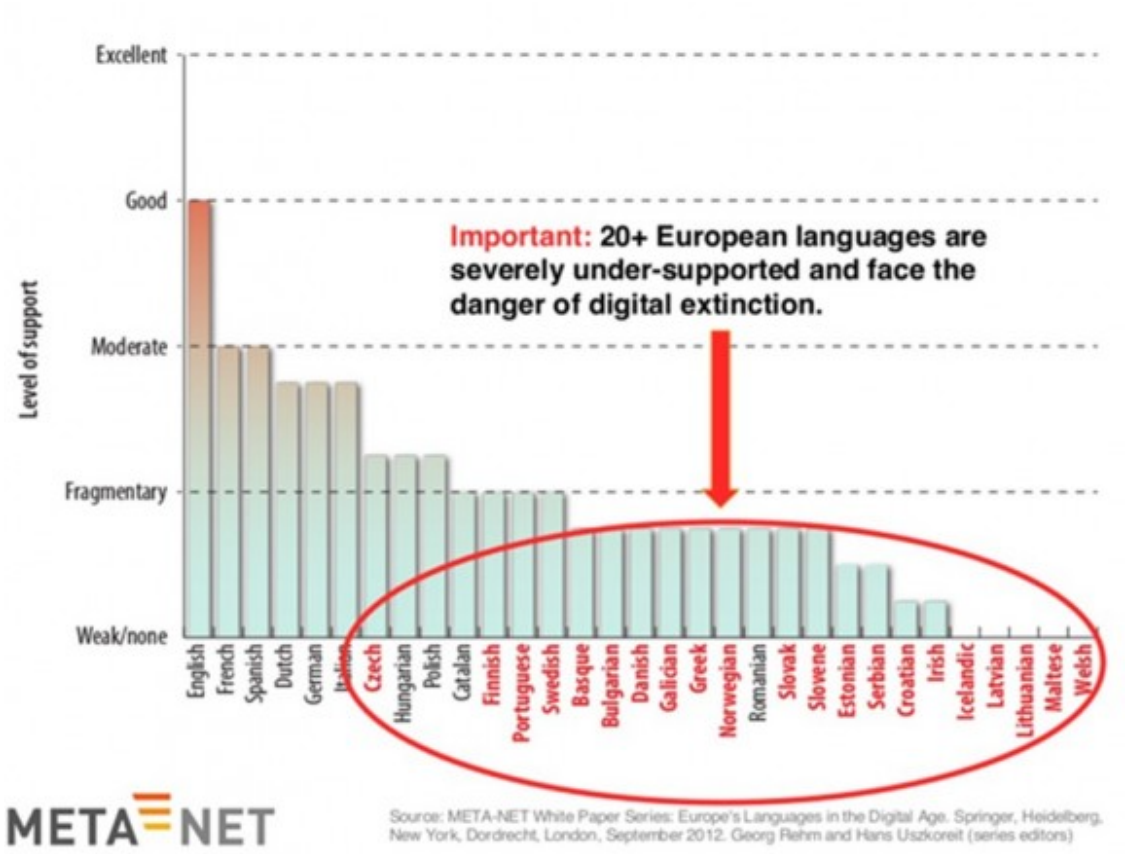
Primer article del dossier «El català al món digital»

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

Hi ha dos aspectes que cal tenir en compte. D'una banda, hem de considerar els recursos disponibles i el suport que té una llengua. De l'altra, hem de considerar l'estat de la tecnologia i l'impacte que té, en el nostre cas, sobre el català.

Pel que fa al suport a la llengua, en general, la societat catalana és una societat força activa que acostuma a mobilitzar-se davant assumptes que considera importants. Hi ha iniciatives i organitzacions que porten molts anys treballant a favor de la llengua, normalment de forma voluntària, amb resultats extraordinàriament bons. Tenim la Viquipèdia, que va ser la tercera en crear-se després de l'anglesa i que actualment és la vintena en nombre total d'articles. Això és extraordinari tenint en compte el nombre de parlants del català. Hem de saber que la Viquipèdia, més enllà de la seva tasca com a divulgació del coneixement, és possiblement el recurs més utilitzat en l'àmbit de les tecnologies de la llengua en qualsevol llengua. **Disposar de prou dades és crític i aquí la Viquipèdia hi juga un paper fonamental**. Tenim també iniciatives con Softcatalà, associació sense ànim de lucre que fomenta l'ús del català a les noves tecnologies i que ha generat un munt de recursos imprescindibles. Aquests són dos exemples evidents i coneguts, però n'hi ha de molts altres que d'una manera o d'altra, des de la societat civil, treballen per a llengua en l'àmbit digital.

Si bé el dinamisme i la implicació de la societat civil és important, no podem ignorar que **la salut del català digital és, també, responsabilitat de l'administració**. En aquest sentit, estem en un punt crític. L'any 2012, un estudi realitzat per més de 200 experts en tecnologies del llenguatge ja advertia que més de 20 llengües europees -i entre elles el català- s'enfrontaven a l'extinció digital si no rebien més suport tecnològic. **Al 2021, la situació és encara pitjor**, per això és fonamental que des de l'administració es faci una aposta decidida i contundent per dotar el català del recursos necessaris per a que podem viure digitalment en català amb la mateixa normalitat que un anglès pot fer-ho en la seva llengua.



Multitud de llengües d'arreu del continent es troben en una situació digital crítica. Font: META-NET White Paper Series: Europe's Language in the Digital Age.

Afortunadament, **l'administració catalana, conscient del problema, ha respost al repte** amb el projecte AINA. L'objectiu d'AINA és generar els recursos 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.

Pel que fa a la tecnologia, la **irrupció de les xarxes neuronals profundes ha suposat una revolució en l'àmbit de la intel·ligència artificial i les tecnologies de la llengua**. Res no és com era només fa uns anys. L'any 2012, desenvolupar recursos per a una llengua era molt costós, mentre que ara la tecnologia permet grans desenvolupaments a menys costos, sempre que disposem de prou dades. Fa un any, hagués dit que la tecnologia jugava al nostre favor: tenint prou dades, costa el mateix entrenar un model per a l'anglès que per al català. Avui, ja no ho tinc tan clar: els *supermodels* com GPT-3 (Brown, 2020) o T5 (Raffel, 2019) són gairebé impossibles per a llengües com la nostra. De fet només n'hi ha per a l'anglès. Hem de pensar que el model GPT-3 d'OpenAI té 175.000 milions de paràmetres, 10 vegades més que qualsevol altre gran model anterior. **Per entrenar-lo van disposar d'un corpus de gairebé 500.000 milions de 'paraules'. El primer model d'AINA s'ha entrenat amb el corpus més gran mai compilat pel català, i té 1.770 milions de paraules**. Si l'aposta és anar cap a *supermodels* com aquests, ho tindrem difícil (com la immensa majoria de les llengües). La majoria de les llengües no tindrà mai un *supermodel* propi, però poden participar en un *supermodel* multilingüe. Els models multilingües com ara mBERT (Devlin, 2018) han demostrat la seva eficàcia, per tant hem d'aconseguir que el català també hi sigui representat. De totes formes, penso que la recerca en models més petits i

eficients i, també, més 'verds' energèticament, serà sempre una necessitat i per tant hem de ser-hi.

Els canvis en la tecnologia ens han ajudat, però recentment hi ha una cursa en veure qui fa el model més gran i amb més dades, i aquí el català -i moltes llengües- patirà. No tenim prou dades per fer un GTP-3 o similar. Tot amb tot, hi ha sortides com els models massius multilingües (emulant el mBERT) i la recerca en obtenir millors models sense tantes dades. La iniciativa BigScience liderada per HuggingFace vol precisament superar l'anglocentrisme i aposta per fer un model massiu multilingüe. Això seria una bona notícia, però en aquesta proposta hi ha vuit llengües i el català no hi és. Haurem d'aconseguir ser-hi, en aquest model o en un altre.

[noticiadiariambautor]93/224[/noticiadiariambautor]
Quines accions s'haurien de promoure perquè fos una llengua disponible a tots els serveis digitals?
S'han de generar els recursos lingüístics necessaris per al català, que permetin i facilitin desenvolupar aplicacions com xatbots, traductors automàtics, assistents de veu, aplicacions d'extracció d'informació, resum automàtics, etc., a costos raonables. La tecnologia de la llengua ha entrat de ple en el món de l'aprenentatge profund i el *big data*, per tant **cal posar a l'abast de la comunitat científica i la indústria models de llengua en català**, pre-entrenats sobre grans quantitats de dades lingüístiques de qualitat. Quan diem dades lingüístiques ens referim a dades multimodals, és a dir dades textuals, dades de veu i de vídeo. Quan diem dades de qualitat, volem dir -també- que han de tenir llicències adequades, han d'estar ben documentades (amb metadades suficients) i han de garantir la traçabilitat.

Cal també generar un bon nombre de dades anotades que serveixin per entrenar i, posteriorment, avaluar models per a tasques específiques (com ara sistemes de pregunta resposta, de classificació semàntica i d'altres tasques que impliquen comprensió del llenguatge). **La generació de dades manualment anotades és un procés crític i molt costós. El català està molt mal dotat en això**. Per tant, cal dedicar-hi recursos suficients.

Cal també, que des de l'administració es duguin a terme les accions necessàries perquè **els grans generadors de dades lingüístiques, com ara la CCMA o la pròpia administració, assegurin el flux de dades** que es necessiten. Això inclou, des de la definició de protocols a seguir per a compartir dades lingüístiques, fins a l'aprovació d'iniciatives legislatives que en garanteixin el subministrament. La reutilització de dades per a usos secundaris (en el nostre cas, usos lingüístics) és una inversió i un estalvi.

Tot això és el que vol resoldre el projecte AINA a través de la producció de la infraestructura necessària agrupada en cinc eixos:

- Desenvolupament de serveis lingüístics bàsics i transversals que serveixin de punt de partida i/o mòduls bàsics sobre els quals desenvolupar aplicacions complexes
- Compilació i preparació de dades massives i de qualitat per a poder entrenar models genèrics de la llengua i models per a tasques específiques, com ara sistemes de pregunta resposta, de diàleg, de resum automàtic, etc.
- Entrenament de models computacionals de llengua, generals i adaptats a domini i/o a tasca, llestos per servir de base per crear noves aplicacions.
- Entrenament de models de reconeixement i síntesi de la parla de qualitat per al català, que puguin ser incorporats als assistents de veu més comuns del mercat
- Entrenament de motors de traducció automàtica entre el català i les principals llengües mundials

Presentació pública del 'projecte AINA', el 10 desembre de 2020. Foto: Generalitat de Catalunya.

Quins actors o recursos caldria activar perquè fos possible?
L'administració, sense cap mena de dubte. Tant pel fet de ser proveïdora de dades com pel fet de ser facilitadora de l'estratègia i els recursos necessaris. I aquí, quan dic recursos necessaris vull dir els recursos econòmics necessaris. **És important també el paper de l'administració com a demandant de tecnologia**, la digitalització de l'administració no només la fa més eficaç i propera al ciutadà, sinó que genera demanda i dinamitza el sector.

Altres actors, són la indústria del sector de la intel·ligència artificial i de la indústria de la llengua: si hi ha demanda i recursos suficients perquè afegir el català a les aplicacions no suposi un cost inassumible, la indústria s'hi posarà. Aquí, quan dic recursos, em refereixo a **recursos lingüístics**: dades, models de la llengua i de la veu, llibreries i mòduls de processament per al català, etc.

Moltes empreses han invertit en els *xatbots* com a eina de comunicació amb els seus clients. Avui en dia, fer un *xatbot* en castellà és assumible, però fer-ho en català és gairebé una heroïcitat. Això és el que hem de solucionar, **no podem normalitzar el fet que quan 'parlem amb una màquina' ho hem de fer en castellà**. Aquí els parlants de la llengua també hi tenim el nostre paper, nosaltres hem de generar la demanda.

El món de la recerca és també un actor essencial. Cal que tingui recursos per avançar en estratègies d'aprenentatge no supervisat per evitar els costos de generar dades anotades; en la generació de models més eficients i més respectuosos energèticament; en mètodes de transferència de coneixement; models multilingües, etc.

[noticiadiariambautor]93/225[/noticiadiariambautor]
Com encaixa la promoció del català en el context d'un mercat global dominat per llengües majoritàries?

Hi ha moltes raons per garantir que la tecnologia estigui disponible en català. Des d'una perspectiva purament empresarial, **s'ha demostrat que la localització d'un producte augmenta la penetració en nous mercats**. Avui en dia la traducció automàtica té una qualitat excel·lent, per tant localitzar (o traduir) al català és perfectament assumible.

La promoció de la llengua pròpia no és una obsessió o raresa del català, és una preocupació de totes les llengües no majoritàries, que són la immensa majoria de les llengües. El setembre del 2018, el Parlament Europeu va aprovar amb una majoria aclaparadora -592 vots a favor i només 45 en contra- la resolució sobre la "Igualtat lingüística a l'era digital" (PDF). Com a resposta a aquesta resolució, va néixer el projecte ELE (European Language Equality -igualtat lingüística europea) l'objectiu del qual és preparar una agenda per a l'estratègica de recerca i innovació i definir un full de ruta per aconseguir la plena igualtat lingüística digital a Europa el 2030.

Com podem veure, **la promoció de les llengües no majoritàries és un objectiu compartit**. No cal justificar-nos.

Com encaixa internament la promoció digital del català en un context de diglòssia a favor del castellà? El català té un doble repte, fer-se lloc en un món digital clarament anglocèntric i encarar els problemes de diglòssia a favor del castellà a casa nostra. És evident que els reptes són importants, però no són del tot nous. No sóc experta en política lingüística, el que sí puc dir és que **les grans llengües digitals tenen molts parlants i molta tecnologia**. És evident que el català no té molts parlants, per tant haurem de tenir una tecnologia abundant i excel·lent.

Bibliografia rellevant:

Brown, T. B., et al. (2020). "Language models are few-shot learners." arXiv preprint arXiv:2005.14165.
Devlin, J., et al. (2018) "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.
EPRS | European Parliamentary Research Service. "Language equality in the digital age"
Melero, M.; R. Costa-jussà, M. (8 d'abril del 2021) AINA, un projecte d'intel·ligència artificial en clau catalana. Blog de la *Revista de Llengua i Dret*.

Moreno, A.; Bel, N.; Revilla, E.; Garcia, E.; Vallverdú, S. (2012). "La llengua catalana a l'era digital - The Catalan Language in the Digital Age". *META-NET White Paper Series*. Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.
Raffel, C., et al. (2019). "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683.
Rehm, G., et al. (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe." arXiv preprint arXiv:2003.13833.
Translating for Europe (8 de novembre de 2018). *Transcending language barriers in the digital age: the EU perspective* [vídeo] https://www.youtube.com/watch?v=sn3QPmaJPhI
Wolf, T., et al. (2020). "Transformers: State-of-the-art natural language processing." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.

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]