REVISTA DE LLENGUA I DRET // JOURNAL OF LANGUAGE AND LAW #70

EL FUTUR DE LES LLENGÜES EN L'ERA DIGITAL: OPORTUNITATS I BRETXA LINGÜÍSTICA

Maite Melero*

Resum

En aquest article reflexionem sobre com impactarà la revolució digital en la supervivència de les llengües en un futur no gaire llunyà. Si una cosa tenim clara és que el llenguatge humà serà el mitjà de comunicació predominant entre les persones i la tecnologia i entre les persones i el coneixement col·lectiu i la informació del món sencer. Efectivament, l'ús d'una llengua o d'una altra determina la quantitat d'informació a la qual es pot accedir, així com els serveis disponibles. La clau és el bagatge tecnològic amb què les diferents llengües s'enfronten al repte digital. La riquesa dels recursos tecnològics de cada llengua afectarà crucialment les seves possibilitats d'arribar amb bona salut al segle XXII. Les llengües en risc més immediat, evidentment, són aquelles afectades per la "diglòssia digital": els parlants bilingües d'una llengua regional i d'una llengua de la globalització, abans que perdre el tren digital, opten per la llengua gran i deixen de banda la que no participa en el progrés tecnològic. Els efectes que això pot tenir en la diversitat lingüística de l'ecosistema digital, i per extensió en el món, són devastadors.

Paraules clau: Tecnologies de la llengua; traducció automàtica; intel·ligència artificial; diversitat lingüística; bretxa lingüística.

THE FUTURE OF LANGUAGES IN THE DIGITAL AGE: OPPORTUNITIES AND LINGUISTIC DIVIDE

Abstract

In this article we reflect on the impact of the digital revolution on the survival of languages in the not too distant future. If one thing is clear, human language will be the predominant means of communication between people and technology, and between people and the collective knowledge and information of the entire world. Indeed, the use of one language or another determines the amount of information that can be accessed, as well as the services available. The key is the technological know-how with which the different languages face the digital challenge. The wealth of technological resources of each language will crucially affect their chances of making it into the 22nd century in good health. The most immediate languages at risk, obviously, are those affected by "digital diglossia": bilingual speakers of a regional language and a global language, rather than missing the digital train, will opt for the larger language and set aside that which does not play a part in technological progress. The effects this may have on the linguistic diversity of the digital ecosystem, and by extension on the world, are devastating.

Keywords: Language technologies; automatic translation; artificial intelligence; linguistic diversity; linguistic divide.

* Maite Melero, doctora especialista en traducció automàtica i processament del llenguatge natural i membre de l'Oficina Tècnica General del Pla d'Impuls de les Tecnologies del Llenguatge a la Secretaria d'Estat per a l'Avenç Digital (SEAD). maite.melero@upf.edu

Citació recomanada: Melero, Maite (2018). El futur de les llengües en l'era digital: oportunitats i bretxa lingüística. *Revista de Llengua i Dret, Journal of Language and Law*, (70), 152-165, DOI: 10.2436/rld.i70.2018.3201.

Sumari

1 Introducció

Les tecnologies que formen part del que es coneix com a "intel·ligència artificial" han experimentat en els últims anys un avenç notable, de tal manera que els canvis que s'esdevindran en un futur proper poden arribar a ser disruptius en molts àmbits. Els assistents personals, en format mòbil o a casa, ja són una realitat quotidiana en les societats més tecnològiques, com ara Estats Units o Singapur. Els cotxes autònoms estan deixant de ser una excentricitat de laboratori per començar a ser considerats més segurs i eficients que els cotxes convencionals. El seu ús massiu en un futur proper és un fet cada cop més probable. La irrupció de les màquines intel·ligents té un impacte social indubtable i genera tota mena de qüestions de tipus legal, laboral i fins i tot ètic, que moltes veus ja s'estan afanyant a plantejar. Tanmateix, de totes les preocupacions que la revolució digital suscita, n'hi ha una que passa molt desapercebuda, fins i tot per als principals afectats: com impactarà la intel·ligència artificial en la supervivència i les possibilitats de desenvolupament i d'ús de les llengües més febles en un futur no gaire llunyà.

L'expansió d'Internet ja representa un repte seriós per a totes les llengües no pertanyents al selecte club de les llengües globals o hegemòniques, que, com veurem, es reparteixen més del 80 % dels continguts a la xarxa. L'ús d'una llengua o d'una altra determina la quantitat d'informació a la qual es pot accedir, així com els serveis disponibles. Tanmateix, amb la irrupció de dispositius domèstics, amb els quals podem dialogar còmodament en llenguatge natural, s'introdueix un nou factor encara més poderós. Ja no és només quina llengua hem d'utilitzar per accedir a informació o serveis en línia, sinó també quina llengua ens permet beneficiar-nos del que promet ser la que ja està considerada com la quarta revolució industrial. Si una cosa tenim clara és que el llenguatge humà serà el mitjà de comunicació predominant entre les persones i la tecnologia i entre les persones i el coneixement col·lectiu i la informació del món sencer. Si la tecnologia pot decidir quin idioma parlem a casa, cal doncs aconseguir que la tecnologia entengui i parli el nostre idioma.

El bagatge tecnològic amb què les diferents llengües s'enfronten al repte digital afectarà les seves possibilitats de desenvolupar-se plenament i de tenir uns usos més amplis en l'entorn digital. La desigualtat entre llengües, en termes tecnològics, posa en un risc real fins i tot aquelles llengües que fins fa poc tenien tot el que calia per garantir la seva continuïtat, és a dir, aquelles considerades com a "segures" en la classificació de la UNESCO (Moseley, 2010). Aquí volem posar de manifest que les llengües que estan més en risc són aquelles afectades per la "diglòssia digital". Els parlants bilingües d'una llengua regional i d'una llengua de la globalització, abans que perdre el tren digital, opten per la llengua gran i deixen de banda la que no participa del progrés tecnològic. Els efectes que això pot tenir en la diversitat lingüística de l'ecosistema digital, i per extensió en el món, són devastadors (Carew *et al*., 2015). Com apunten informes recents del Parlament Europeu, calen polítiques adequades, junt amb l'apoderament de les mateixes comunitats lingüístiques, per afrontar la nova situació creada per la revolució digital i aconseguir que els beneficis que ens ha d'aportar arribin en igualtat de condicions a totes les persones, independentment de quina llengua parlin.

2 La revolució de les xarxes neuronals i l'aprenentatge profund

L'any 2012, l'algorisme d'aprenentatge profund o *deep learning* de tres investigadors de la Universitat de Toronto va quedar el primer en la competició anual de reconeixement d'imatges ImageNet,[1] ja que va obtenir uns resultats sorprenents, que eren un 41 % millors que el segon classificat. L'algorisme en qüestió, una xarxa neuronal convolucional, era conegut des de feia dècades (Krizhevsky et *al*., 2017), però les noves condicions en termes de potència computacional i sobretot en quantitat de dades per a l'aprenentatge li van donar una segona oportunitat i van iniciar la revolució en intel·ligència artificial que promet canviar les nostres vides. A partir de la publicació dels resultats de la competició, investigadors en reconeixement d'imatge d'arreu van començar a utilitzar aquest tipus d'algorismes i els resultats en la tasca de reconeixement immediatament van experimentar una millora sense precedents, fins al punt que, actualment, la competició ja no existeix perquè es considera que la tasca que proposava ImageNet ja està resolta. La taxa d'error actual és increïblement baixa, al voltant del 2 %. Aplicant algorismes semblants, avui en dia, Google Photos ens permet cercar objectes o persones en els nostres propis àlbums de fotos al núvol, en temps real.

1 http://www.image-net.org/ [Consulta: 28 juliol 2018]

Com sol passar en ciència, els avenços en un camp no triguen a impactar en d'altres. Així, moltes altres àrees de classificació automàtica, que comptaven amb quantitats prou grans de dades, van començar a experimentar amb algorismes d'aprenentatge profund i van ampliar els èxits obtinguts en reconeixement d'imatge a altres àmbits de la intel·ligència artificial, com el processament automàtic del llenguatge natural. Actualment, gràcies a l'aprenentatge profund, s'està progressant de forma espectacular en tasques molt complexes, com ara el reconeixement de veu independent del locutor, la gestió del diàleg i el raonament, i la traducció automàtica.

Contínuament estan apareixent aplicacions i productes intel·ligents basats en l'aprenentatge profund sobre grans quantitats de dades, en una gran diversitat d'àrees, algunes de gran impacte social. Estem assistint als inicis d'una revolució digital que està començant a transformar-ho tot, des de les comunicacions i la informàtica fins a la medicina, la fabricació i el transport. Els dispositius intel·ligents estan entrant també en els entorns domèstics, amb la domòtica, la Internet de les coses, els robots de servei, els assistents virtuals i la conducció autònoma. En aquests entorns, el mitjà de preferència per interactuar amb la màquina és i serà cada cop més la veu humana, el diàleg parlat, pels seus avantatges evidents: mans lliures, distància física, coneixement previ del codi, etc.

Ja fa anys que els assistents virtuals ens permeten posar-nos música, afegir cites a l'agenda, fer trucades o enviar missatges, però el sistema anomenat Duplex,[2] que Google va presentar el passat mes de maig, suposa un salt de gegant sobre el que ja coneixem. Duplex no només ens posa la trucada quan li ho demanem, com ja fa l'actual assistent de Google, sinó que gestiona ell mateix tota la conversa. A la presentació esmentada, es demostrava com el sistema artificial era capaç de fer una reserva de taula en un restaurant, amb un diàleg fluid i natural, amb gran precisió en les interaccions parlades i gran naturalitat en la fonètica i prosòdia, fins i tot amb petites disfluències com ara "ee" o "mmm". Duplex sona tan natural que una de les crítiques d'aquella primera demostració era que podia induir a engany al seu interlocutor, de manera que en la segona demostració pública, ocorreguda un parell de mesos més tard, s'inicia la conversa amb l'avís a l'interlocutor que està parlant amb un sistema artificial. Una de les coses més impressionants és que no es tracta d'una única "veu Duplex", o d'un conjunt limitat de veus, sinó que cada trucada es pot fer amb una veu amb personalitat pròpia (femenina, masculina, jove, nasal, madura, etc.). Les variacions potencials són infinites. WaveNet,[3] el model que genera aquestes veus, desenvolupat al departament de Deepmind de Google, és capaç de general qualsevol tipus de so bucal humà, sospirs, espetecs, etc. Es tracta d'un model de xarxes neuronals basat en ones acústiques, que supera les limitacions dels sistemes anteriors de generació de veu, tant els que concatenen segments gravats prèviament, com els paramètrics, que són més flexibles però que sonen menys naturals. WaveNet és extraordinàriament flexible i natural. Com que treballa directament sobre l'ona acústica pot reproduir qualsevol tipus de so, per exemple música, o falsos idiomes. De fet, per ser capaç de generar llengües reals, el model acústic ha d'acompanyar-se també d'informació lingüística (fonètica, sil·làbica, etc.).

La qualitat de la veu no és l'únic aspecte sorprenent de Duplex. La flexibilitat i intel·ligència del seu diàleg ha fet afirmar als seus creadors que Duplex seria el primer sistema artificial en passar el test de Turing, si més no, en la tasca per la qual ha estat entrenat, és a dir concertar una cita per telèfon (al restaurant, perruquer, etc.). El test de Turing és la prova proposada el 1950 per l'informàtic anglès Alan Turing com una forma d'avaluar la intel·ligència d'una màquina: per superar-la un robot ha de comportar-se d'una manera indistingible d'un ésser humà. Tanmateix, les capacitats actuals de Duplex són només el principi del que es pot arribar a aconseguir. Hi ha infinitat de tasques per a les quals aquest auxiliar telefònic podria ser útil, tant per al client que necessita concertar un servei, com de la de l'empresa subministradora del servei. Arribarà un moment —i no falta pas gaire— en què els robots mantindran converses entre ells i gestionaran de forma satisfactòria aquestes i altres tasques.

2 https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html [Consulta: 28 juliol 2018]

3 https://deepmind.com/blog/wavenet-generative-model-raw-audio/ [Consulta: 28 juliol 2018]

3 La irrupció de les màquines intel·ligents en les nostres vides

Stephen Hawking ha estat una de les primeres veus importants d'advertir dels perills potencials de la intel·ligència artificial. En una xerrada al Web Summit de 2017 a Lisboa,[4] pocs mesos abans de la seva mort, el prestigiós físic va afirmar que la intel·ligència artificial seria o bé "el millor" o bé "el pitjor" que li podria passar a la humanitat, depenent de si la societat és capaç o no de mantenir el control sobre el seu desenvolupament. Hawking va afegir que "els ordinadors poden, en teoria, emular la intel·ligència humana, i superar-la" i va ressaltar el seu potencial positiu per a afrontar reptes mundials com la pobresa i la malaltia. Però també va prevenir de grans riscos, com la disrupció que poden tenir sobre l'economia i el mercat del treball, la capacitat destructiva de les armes autònomes i, en definitiva, "noves maneres d'opressió d'uns quants sobre la majoria".

Efectivament, la intel·ligència artificial ja iguala o supera el rendiment humà en un nombre creixent de sectors. Diverses tasques tradicionalment realitzades per humans ja han estat assumides per robots i algorismes. A més, el consens general és que les capacitats de la intel·ligència artificial seguiran creixent i el seu ús es generalitzarà ràpidament. El creixement exponencial de les capacitats i l'aplicabilitat de la intel·ligència artificial ha suscitat preocupació per l'automatització de l'ocupació i la possibilitat d'un atur tecnològic massiu, però també pel seu impacte a la baixa en els salaris dels treballadors que corren més risc de ser desplaçats. S'està constatant una major polarització dels mercats laborals, amb un augment de la proporció de llocs de treball d'alta i baixa qualificació, d'una banda, i una forta disminució de la proporció de llocs de treball rutinaris de qualificació mitjana.[5]

En un article publicat el maig passat a Modern Diplomacy[6] es reflexiona sobre les dificultats ètiques i legals plantejades per uns sistemes que són cada vegada més autònoms a l'hora de realitzar tasques complexes i per la disminució de la capacitat dels humans per comprendre, predir i controlar-ne el funcionament. Molts d'aquests sistemes tenen la capacitat d'aprendre de la seva pròpia experiència i dur a terme accions més enllà de l'abast de les previstes pels seus creadors. Tanmateix, la predicibilitat és un factor crucial a les legislacions modernes (Asaro, 2016). Aquest tipus de dificultats eticolegals són les que estan frenant parcialment l'expansió dels cotxes autònoms. En els cursos d'ètica de les facultats, s'explica un conegut dilema moral: un tramvia baixa descontrolat per una via en la qual es troben lligades cinc persones. Hi ha temps d'accionar una palanca que desviarà el tramvia a una altra via on hi ha una persona lligada. Accionaríem la palanca? En el cas dels cotxes autònoms, és fàcil imaginar situacions en què un accident sigui inevitable, i el cotxe hagi de prioritzar la vida dels passatgers, dels vianants o de cap dels dos. Per investigar les reaccions humanes en aquestes situacions, l'Institut Tecnològic de Massachusetts (MIT) ha creat un lloc web on es presenten escenaris d'aquesta mena i es demana a l'usuari prendre decisions.[7]

Els algorismes d'aprenentatge automàtic estan sotmesos inevitablement al biaix que presenten les dades sobre les quals s'entrenen. Aquest és un altre problema greu, que ja es va posar de manifest als Estats Units el 2016, amb l'apel·lació d'un home condemnat a una llarga pena de presó que tenia com a base la informació obtinguda mitjançant un algorisme que calculava la probabilitat que reincidís en el delicte.[8] Alguns governs ja estan començant a plantejar-se aquestes qüestions. A finals de març del 2018, el president de França, Emmanuel Macron, va presentar la nova estratègia nacional d'intel·ligència artificial del país, amb una inversió de 1.500 milions d'euros en els pròxims cinc anys per donar suport a la recerca i la innovació en aquest camp, enfocada a quatre sectors específics: l'assistència sanitària, el transport, el medi ambient i la seguretat. L'estratègia francesa proposa un seguit de mesures, com ara desenvolupar algorismes transparents, que puguin ser provats i verificats, determinar la responsabilitat ètica dels qui treballen en intel·ligència artificial i crear un comitè d'assessorament d'ètica. Per la seva banda, la Unió Europea, en la seva resolució sobre el Reglament de Dret Civil sobre Robòtica, proposa dos codis de conducta per tractar les qüestions

4 https://www.evolving-science.com/intelligent-machines-artificial-intelligence/stephen-hawking-web-summit-will-artificial-intelligence-help-us-or-destroy-us-00472

5 http://www.oecd.org/els/emp/future-of-work/artificial-intelligence-and-the-labour-market-should-we-be-worried-or-excited.htm [Consulta: 28 juliol 2018]

6 https://moderndiplomacy.eu/2018/04/24/the-ethical-and-legal-issues-of-artificial-intelligence/ [Consulta: 28 juliol 2018]

7 http://moralmachine.mit.edu/ [Consulta: 28 juliol 2018]

8 https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html [Consulta: 28 juliol 2018]

ètiques:[9] un Codi de Conducta Ètica per als Enginyers en Robòtica i un Codi per als Comitès d'Ètica de la Investigació. El primer codi proposa quatre principis ètics en l'enginyeria robòtica: 1) beneficència (els robots han d'actuar en el millor interès dels humans); 2) no maleficència (els robots no han de danyar els humans); 3) autonomia (la interacció humana amb els robots ha de ser voluntària); i 4) justícia (els beneficis de la robòtica s'han de distribuir equitativament).

Totes aquestes qüestions, i moltes d'altres, particularment les que afecten la disrupció socioeconòmica, són d'importància capital i estan començant a formar part de l'agenda política i els programes de discussió dels dirigents i agents socials i econòmics. En el que queda d'article volem cridar l'atenció sobre un aspecte que no sol formar part d'aquestes agendes i discussions i que tanmateix també té un gran impacte sobre els drets d'una part substancial de la població mundial, incloent-hi l'europea, i en particular la catalana. Com hem apuntat més amunt, el llenguatge humà serà el mitjà de comunicació predominant entre les persones i la tecnologia i entre les persones i el coneixement col·lectiu i la informació del món sencer. Si considerem que tots els ciutadans tenen dret a accedir a la tecnologia i al coneixement col·lectiu en les mateixes condicions, el que està en risc és que puguin fer-ho tots en la llengua pròpia.

4 La bretxa lingüística a l'era digital

En un món hiperconnectat, la supervivència de moltes de les llengües actuals està compromesa. Sutherland, en un conegut article a *Nature* (Sutherland, 2003) on fa un paral·lelisme entre l'extinció de llengües i espècies, afirma que la desaparició de les llengües va a un ritme més ràpid que la de les espècies. En aquest article, apunta com a causes de la pèrdua de la diversitat lingüística a factors socials o econòmics (migració, globalització del comerç i dels mitjans de comunicació), però també a polítiques nacionals desfavorables i a la diferència de prestigi que es genera entre les llengües dominants i les que no ho són. En canvi, molt més rarament la desaparició d'una llengua s'associa amb fenòmens naturals, com l'extinció d'una població. A l'article "Digital Language Death" (Kornai, 2013), András Kornai assenyala que si bé és cert que una llengua no desapareix fins que en mor l'últim parlant, hi ha tres senyals que anticipen aquesta desaparició: pèrdua de funció, pèrdua de prestigi i pèrdua de competència. Si hem de creure allò que "el que no és a la Web, no existeix", les llengües sense presència suficient a Internet, no "serveixen", no tenen prestigi i no tenen "nadius digitals" que les facin servir. Segons Kornai, fins al 95 % de les llengües que avui encara es parlen al món presenten aquests senyals de propera desaparició. Una mort digital massiva.

Internet va començar com una xarxa gairebé monolingüe: cap el 1998 el 80 % dels llocs web eren en anglès. Actualment, altres idiomes grans han entrat amb força: xinès, espanyol, àrab, portuguès i malai segueixen l'anglès, que ja *només* representa el 53,1 % del contingut a la xarxa.[10] Japonès, rus, francès i alemany completen els 10 grans. Aparentment Internet es va fent cada vegada més multilingüe i culturalment divers, tot i que el 80 % dels continguts està en alguna de les 10 llengües grans, mentre que la resta es reparteixen el 20 % restant.

L'ús d'una llengua o una altra determina necessàriament l'experiència de l'internauta, així com la quantitat d'informació a la qual pot accedir, i potser també la qualitat d'aquesta informació, o els serveis que té disponibles, les comunitats amb les quals pot relacionar-se, etc. Internet pot semblar infinit, però només és tan gran com la llengua que fem servir per moure'ns per la xarxa. Parafrasejant Wittgenstein, els límits del meu idioma són els límits del meu món en línia. Fins i tot un parlant d'una llengua majoritària té una visió limitada de la diversitat d'informació disponible. Per posar un exemple, un usuari que només parli anglès —el Goliat de les llengües— no podrà accedir a la meitat d'articles de la Viquipèdia alemanya (la segona més gran)[11] perquè no tenen correspondència a la Viquipèdia anglesa, ni tampoc al 70 % dels articles de la Viquipèdia en txec o en italià, per la mateixa raó.

La bretxa digital global es tanca quan les capacitats de connectivitat i accés a la xarxa van arribant a tots els territoris i capes socials. És cert que a mesura que es tanca la bretxa digital creixen comunitats lingüístiques minoritàries a Internet i se n'incorporen de noves, però això pot ser només un miratge. És un fet que l'accés

9 http://www.europarl.europa.eu/RegData/etudes/STUD/2016/571379/IPOL_STU(2016)571379_EN.pdf [Consulta: 28 juliol 2018]

10 https://w3techs.com/technologies/overview/content_language/all [Consulta: 28 juliol 2018]

11 https://stats.wikimedia.org/EN/TablesArticlesTotal.htm [Consulta: 28 juliol 2018]

a Internet ofereix oportunitats per a l'apoderament lingüístic de comunitats parlants de llengües minoritàries, i és una eina molt potent per documentar i preservar llengües en risc. Fins i tot trobem comunitats virtuals a Internet de llengües que ja són mortes en el món real: el projecte Muysccubun, per exemple, ha estat treballant per documentar i compartir l'extint idioma Muisca, parlat antigament al centre de Colòmbia, creant diccionaris en línia i construint una comunitat al voltant de seva pàgina de Facebook.[12] Tanmateix, l'accés a Internet de comunitats amb una llengua minoritària, en un context d'una altra llengua forta, pot tenir l'efecte contrari i arribar a accelerar la desaparició de la llengua minoritària. Això està passant en les comunitats aborígens australianes, on, com més es tanca la bretxa digital, més creix la bretxa lingüística generacional, amb una generació jove sense competències lingüístiques en la llengua pròpia, en benefici de l'anglès (Carew, 2015).

La situació actual, doncs, es pot resumir de la manera següent: Internet creix i creix també en diversitat lingüística, però un conjunt reduït de llengües són molt més utilitzades que la resta, amb clara predominança de l'anglès. Aquesta segueix sent, de lluny, la llengua més utilitzada a Internet, la llengua per a la qual es generen més continguts i també la llengua privilegiada per a la majoria dels usuaris. És natural que, en incorporar-se al món digital, l'individu adopti les llengües que li donen accés a més continguts, però aquesta tendència es realimenta i fa créixer la bretxa digital entre les llengües. Així, mentre disminueix la bretxa digital entre territoris, perquè tots es van incorporant a la digitalització, podria estar augmentant la bretxa entre les llengües. Claudia Soria, directora del Projecte de Diversitat de Llengües Digitals,[13] que impulsa la diversitat lingüística a Internet, ens avisa de les conseqüències d'aquesta bretxa: la quantitat de productes i serveis que estan disponibles en les llengües menys parlades es redueix proporcionalment, i es crea així desigualtat a diferents nivells (Linguapax Review, 2016):

- Desigualtat de drets lingüístics i oportunitats digitals per a totes les llengües i per a tots els ciutadans. Per exemple, el traductor de Google no inclou l'occità.
- Desigualtat d'accés a la informació i als serveis. Per exemple, la Viquipèdia en anglès té 5,7 milions d'articles, la segona (alemanya) només en té la meitat; Facebook només dona suport a 147 llengües, Booking a 43 i TripAdvisor a 29.
- Accés desigual al desenvolupament tecnològic.
- Desigualtat d'oportunitats per a la supervivència lingüística.

5 Les tecnologies de la llengua

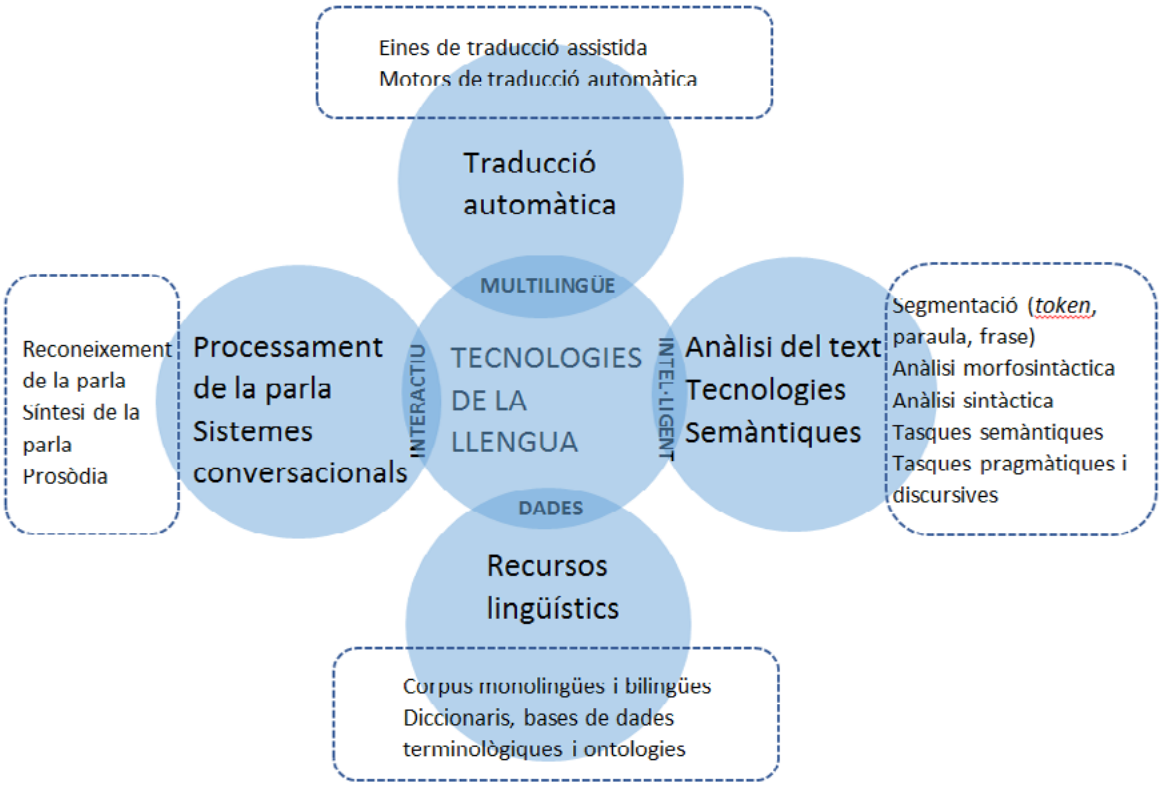
Tots els assistent virtuals, domèstics i mòbils actuals, com Alexa, Siri, Cortana, Google Now, Google Home i ara Duplex, ja són capaços de reconèixer la veu humana, d'extreure significat de comandes i preguntes en llenguatge natural, de raonar sobre elles i d'emetre respostes en veu sintètica. Les comunicacions mòbils, els mitjans socials, els assistents intel·ligents i les interfícies basades en veu estan transformant la forma en què els ciutadans, les empreses i les administracions públiques interactuen i es relacionen. Els impressionants avenços aconseguits en els últims anys ens permeten preveure un futur molt proper en el qual la interacció natural amb les aplicacions intel·ligents serà la norma. Darrere de tots aquests productes que utilitzen el llenguatge d'una manera o altra trobem el que anomenem "tecnologies de la llengua".

Però la utilitat de les tecnologies de la llengua va molt més enllà que dotar de capacitats de diàleg els assistents virtuals. L'anomenada "mineria de textos", o *text mining*, facilita l'extracció d'informació a partir de quantitats enormes de dades i permet estructurar aquesta informació, de manera que pugui ser interpretada de forma automàtica, agregada, classificada, enllaçada i visualitzada. La necessitat d'automatitzar l'anàlisi de gran quantitat de dades afecta àrees tan diverses com la sanitat (p. e. anàlisi d'informes clínics per ajudar a diagnosticar malalties o descobrir interaccions entre fàrmacs), la justícia (p. e. anàlisi de jurisprudència), les finances (p. e. prediccions de moviments borsaris), el màrqueting (p. e. seguiment de marques o campanyes a les xarxes socials) o la vigilància sectorial (p. e. detecció de plagis, duplicació de patents, etc.).

12 https://rising.globalvoices.org/blog/2015/05/09/bringing-the-muisca-language-back-from-extinction-one-word-at-a-time/ [Consulta: 28 juliol 2018]

13 http://www.dldp.eu/ [Consulta: 28 juliol 2018]

Les tecnologies de la llengua són programes dissenyats per analitzar i gestionar el llenguatge humà, en totes les seves formes: parlat, escrit i fins i tot signat. Comprenen un grup gran i heterogeni de tècniques i eines bàsiques per a l'anàlisi i la producció del llenguatge. La il·lustració de sota presenta una visió general d'aquestes tecnologies i tasques representatives:



Font: Adaptat de l'Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información

Els algorismes d'aprenentatge profund també han suposat una revolució per aquestes tecnologies: han millorat espectacularment els resultats obtinguts en totes les àrees, des del reconeixement de veu fins la traducció automàtica, passant per l'anàlisi semàntica. La majoria de les eines de processament del llenguatge, incloent-hi la traducció automàtica, actualment es basen en algorismes de xarxes neuronals, en detriment d'altres estratègies estadístiques o de coneixement lingüístic formalitzat en forma de regles.

En conseqüència, els conjunts de dades o recursos lingüístics específics sobre els quals s'entrenen els algorismes adquireix gran importància. Per posar un exemple, els sistemes de traducció automàtica necessiten grans col·leccions de text bilingüe alineat. Els reconeixedors de veu necessiten corpus d'àudio acompanyats de transcripcions. Altres eines de processament necessiten corpus de text amb anotacions específiques per a la tasca per a la qual s'han d'entrenar, així com recursos lingüístics estructurats com ara gramàtiques, lèxics, tesaurus, terminologies, diccionaris i ontologies.

6 Les dades són el nou petroli

Clive Humby, matemàtic britànic i arquitecte de Tesco's Clubcard, va declarar el 2006: "Les dades són el nou petroli. Són valuoses, però si no es refinen no es poden utilitzar a la pràctica. [Igual que el petroli] cal transformar-lo en gas, plàstic, productes químics, etc. per crear un producte útil i rendible; també cal descompondre i analitzar les dades perquè tinguin valor".[14] Les tecnologies de la llengua són una tecnologia clau per a l'aprofitament intel·ligent de quantitats massives de dades.

14 https://www.theguardian.com/technology/2013/aug/23/tech-giants-data [Consulta: 28 juliol 2018]

Els tres grans pilars sobre els quals se sosté la intel·ligència artificial són: els algorismes, la potència de càlcul i, de manera fonamental, les dades. No és casualitat que els avenços en intel·ligència artificial estiguin liderats per grans empreses com Google, Amazon, Facebook i Microsoft, que maneguen enormes quantitats de dades. Segons David Kenny, director general del servei d'intel·ligència artificial Watson d'IBM, només el 20 % de la informació mundial s'emmagatzema a Internet, mentre que l'altre 80 % està en mans d'empreses i organitzacions privades.[15] El progrés tecnològic depèn de la capacitat d'accedir a grans quantitats de dades i recursos lingüístics de qualitat. La manca d'accés a aquestes dades limita molt significativament el desenvolupament d'aquestes tecnologies. Un accés convenient i ben regulat a les dades és imprescindible per al desenvolupament de nous productes, aplicacions i serveis. Les polítiques de dades obertes són essencials per a la innovació en intel·ligència artificial i processament del llenguatge. Els models de negoci de les empreses generadores de dades i unes polítiques reguladores insuficients provoquen el control de les dades a mans d'un conjunt reduït d'agents i limiten greument la investigació i el desenvolupament tecnològic. Cal promoure polítiques adequades de dades obertes basades en l'ètica, la transparència i l'accessibilitat a les dades, tant del sector privat com del sector públic, tot garantint els drets de la ciutadania a la privacitat i la confidencialitat. I cal també una estreta cooperació entre la indústria i els diferents organismes que generen, posseeixen, necessiten i utilitzen les dades.

Per tal que una llengua pugui participar en la revolució digital, cal que estigui ben dotada tecnològicament. Dues coses són essencials: un conjunt suficient de recursos lingüístics anotats (corpus, lèxics, ontologies) i un accés a grans quantitats de dades en aquesta llengua.

7 Multilingüisme i traducció automàtica

La necessitat de comunicació entre comunitats lingüístiques diverses i d'accés a quantitats més grans de continguts i de serveis ha comportat l'èxit de l'anglès com a *lingua franca*. Tanmateix, les tecnologies de la llengua i en particular la traducció automàtica són claus per fer desaparèixer les barreres idiomàtiques i, en conseqüència, permetre la interacció multilingüe i fer innecessària l'existència d'una o més *lingua franca*. Els avenços tecnològics estan convertint la traducció automàtica en una solució real per superar aquestes barreres. La creixent disponibilitat de serveis de traducció d'alta qualitat està creant l'entorn tecnològic adequat perquè empreses i organismes públics tinguin l'oportunitat d'oferir els seus serveis i continguts directament en l'idioma de l'usuari.

La idea d'utilitzar ordinadors per traduir llengües naturals es remunta al 1950, en el context de la Guerra Freda. La metodologia del moment era molt bàsica: consistia a utilitzar diccionaris bilingües per fer traduccions paraula a paraula. Evidentment, el mètode resultava extraordinàriament limitat a causa de l'ambigüitat, la polisèmia i les diferències en l'ordre de les paraules i en les gramàtiques de les llengües. Els sistemes que van venir després, basats en gramàtiques computacionals programades per lingüistes experts, van millorar els resultats, però suposaven un procés manual, llarg i costós i seguien sense gestionar bé l'ambigüitat i complexitat del llenguatge. Entre 1990 i principis del 2000, a mesura que la potència computacional i l'accés a les dades augmentaven i s'abaratien, va créixer l'interès pels enfocaments estadístics. Els sistemes estadístics de traducció automàtica extreuen automàticament les regles de traducció a partir del text ja traduït per humans, que s'anomena corpus bilingüe o paral·lel, en què cada frase o segment s'alinea amb la seva traducció corresponent en l'altre idioma. El corpus multilingüe de les Nacions Unides i l'Europarl, que conté les actes del Parlament Europeu en 11 llengües, són importants corpus paral·lels que s'han utilitzat sovint per entrenar sistemes de traducció. Els sistemes estadístics aprenen de les dades, requereixen menys esforç humà i són capaços de captar particularitats de l'idioma (per exemple, expressions idiomàtiques) més difícils de tractar en els sistemes basats en el coneixement reglat. El 2012, un sistema estadístic gratuït finançat per la Comissió Europea anomenat Moses[16] va esdevenir la base sobre la qual s'han construït multitud de sistemes de traducció comercials. Al mercat professional, les agències de traducció, algunes de les quals també estan desenvolupant activament els seus propis motors de traducció utilitzant programari de codi obert com Moses, estan passant progressivament de la "traducció humana" a la "traducció automàtica amb

15 http://fortune.com/2016/07/11/data-oil-brainstorm-tech/ [Consulta: 28 juliol 2018]

16 http://www.statmt.org/moses/ [Consulta: 28 juliol 2018]

postedició humana". El cost de la traducció humana sense automatització és simplement insostenible en la majoria dels casos.

Cap al 2014, Joshua Bengio, de la Universitat de Mont-real, va aplicar l'aprenentatge profund a la traducció automàtica (Cho *et al*., 2014) i va començar la substitució del paradigma estadístic pel de les xarxes neuronals. El gran avenç de la traducció automàtica neuronal és que el context de traducció passa de ser un segment d'unes quantes paraules a tota la frase. El resultat és una traducció de qualitat quasi-humana, si es disposen de les dades d'entrenament suficients. Actualment, pràcticament tots els motors de traducció que s'estan desenvolupant, ja sigui a les grans corporacions (Google, Microsoft) o petites companyies, són de base neuronal.

8 Desigualtat tecnològica entre les llengües: la situació del català

El 2012 es va publicar la Sèrie de Llibres Blancs META-NET "Les llengües d'Europa en l'era digital" (Rehm i Uszkoreit, 2012), un estudi sistemàtic de les característiques de 30 llengües europees i del grau de suport tecnològic per cadascuna d'elles. L'estudi, emprès per més de 200 experts i documentat en 31 volums, valorava el suport tecnològic en quatre àrees diferents: anàlisi automàtica de textos, traducció automàtica, processament de la parla i disponibilitat de recursos lingüístics.

Les 30 llengües europees incloses a l'estudi eren: alemany, anglès, basc, búlgar, català, croat, danès, eslovac, eslovè, espanyol, estonià, finès, francès, gal·lès, gallec, grec, holandès, hongarès, irlandès, islandès, italià, letó, lituà, maltès, noruec, polonès, portuguès, romanès, serbi, suec i txec. Les quatre àrees es van avaluar segons els següents paràmetres:

● Traducció automàtica: qualitat de les tecnologies existents; nombre de parells d'idiomes coberts; cobertura de fenòmens lingüístics i àmbits temàtics; aplicacions de traducció automàtica disponibles.

● Processament de la parla: qualitat de les tecnologies de reconeixement i síntesi de veu existents; cobertura dels diferents àmbits temàtics; aplicacions disponibles.

● Anàlisi automàtica de textos: qualitat i cobertura de les tecnologies existents (morfologia, sintaxi, semàntica); cobertura de fenòmens lingüístics i àmbits temàtics; quantitat i varietat d'aplicacions disponibles.

● Recursos: nombre, qualitat i mida de corpus monolingües; nombre, qualitat i mida de corpus paral·lels; nombre, cobertura, qualitat i mida de diccionaris, gramàtiques, glossaris, ontologies i recursos lèxics.

La taula de sota resumeix els resultat de l'anàlisi per a les 30 llengües en les quatre àrees avaluades, segons una escala de quatre nivells: bon suport, suport moderat, suport fragmentari i suport inexistent o feble.

Tecnologia	Bon suport	Suport moderat	Suport fragmentari	Suport feble o inexistent
Traducció automàtica	anglès	francès, espanyol	català, holandès, alemany, hongarès, italià, polonès, romanès	basc, búlgar, croat, txec, danès, estonià, finès, gallec, grec, islandès, irlandès, letó, lituà, maltès, noruec, portuguès, serbi, eslovac, eslovè, suec, gal·lès

Processament de la parla	anglès	txec, holandès, finès, francès, alemany, italià, portuguès, espanyol	basc, búlgar, català, danès, estonià, gallec, grec, hongarès, irlandès, noruec, polonès, serbi, eslovac, eslovè, suec	croat, islandès, letó, lituà, maltès, romanès, gal·lès
Anàlisi automàtica de textos	anglès	holandès, francès, alemany, italià, espanyol	basc, búlgar, català, txec, danès, finès, gallec, grec, hongarès, noruec, polonès, portuguès, romanès, eslovac, eslovè, suec	croat, estonià, islandès, irlandès, letó, lituà, maltès, serbi, gal·lès
Recursos lingüístics	anglès	txec, holandès, francès, alemany, hongarès, italià, polonès, espanyol, suec	basc, búlgar, català, croat, danès, estonià, finès, gallec, grec, noruec, portuguès, romanès, serbi, eslovac, eslovè	islandès, irlandès, letó, lituà, maltès, gal·lès

Font: Extret dels resultats de la sèrie de Llibres Blancs de META-NET

El 2014 es va dur a terme una ampliació de l'estudi (Rehm *et al*., 2014) per tal d'incloure les llengües regionals i minoritàries europees de més de 100.000 parlants que havien quedat fora en la primera fase, com l'albanès, el bretó o el bable. La comparació creuada de les dades obtingudes mostra l'abisme tecnològic entre les diferents llengües europees. Així, hi ha un conjunt reduït de llengües, en general majoritàries, que disposa d'eines i recursos de bona qualitat per a certes àrees d'aplicació, mentre que la resta estan clarament infradotades, en alguns casos de forma severa. Com s'aprecia a la taula superior, el suport digital per a 21 dels 30 idiomes investigats és "inexistent", o "feble" en el millor dels casos. L'altra dada que destaca és la posició privilegiada de l'anglès en el conjunt de les llengües; comparades amb l'anglès, totes les llengües sense excepció queden enrere tecnològicament. Sis anys després d'aquest estudi, la bretxa tecnològica entre l'anglès i les altres llengües europees no ha parat de créixer (Rivera *et al*., 2017). No només hi ha més incentius comercials per desenvolupar solucions per a l'anglès, sinó que, en una mena de cercle viciós, l'existència d'eines i recursos ja disponibles per a l'anglès fa que sigui fàcil provar noves idees per a aquesta llengua, mentre que començar a explorar en altres llengües requereix un cost inicial més alt en termes de models bàsics, de manera que els investigadors són menys propensos a treballar-hi.

L'estiu del 2015 es va presentar un informe sobre l'estat de les tecnologies de la llengua a Espanya (Bel i Rigau, 2015), actualitzat el maig del 2018. L'informe feia una anàlisi FODA (fortaleses, debilitats, oportunitats i amenaces) sobre la disponibilitat de recursos lingüístics i la situació de la recerca i la indústria de la llengua a Espanya, i també ressaltava els usos potencials d'aquestes tecnologies per a l'Administració pública. Un any després, Núria Bel, una de les autores d'aquest informe, va enfocar l'anàlisi a Catalunya i a la situació del català en la publicació "Les indústries de la llengua i la tecnologia per al català" (Bel i Marimon, 2016). En aquest document es ressaltava com a fortalesa l'interès precoç que aquestes tecnologies van despertar a Catalunya, particularment la traducció automàtica, que ja des dels anys noranta del segle passat es va utilitzar, primer en publicacions bilingües de premsa (*El Periódico*, *Segre* i, més tard, *La Vanguardia*), i després en l'Administració de justícia. El clúster català d'indústries de la llengua, ClusterLingua, creat el 2011, engloba una vintena d'empreses del sector, algunes amb presència internacional, en particular en el

reconeixement de la parla i la traducció automàtica. També la desena de grups de recerca dedicats a aquestes tecnologies, repartits a totes les universitats catalanes, compten majoritàriament amb projectes de recerca i innovació de finançament europeu, nacional o de col·laboració amb empreses. D'entre els productes desenvolupats destaca el processador FreeLing (Padró i Stanilovsky, 2012), programa de codi obert d'anàlisi lingüística del català i altres llengües, amb més de 250.000 descàrregues des del 2009. La majoria de recursos per al processament del català recollits a l'informe han estat produïts amb finançament públic pels grups d'investigació de les diferents universitats catalanes, a més de l'Institut d'Estudis Catalans i del TERMCAT, el centre de terminologia per a la llengua catalana.

En aquest informe es considera una debilitat crucial la manca de visibilitat, a nivell mundial, que pateix el sector de les tecnologies lingüístiques, l'important paper de les quals també s'ignora quan es parla d'intel·ligència artificial. Aquesta ignorància s'estén a la manca de conscienciació a l'hora de valorar els recursos lingüístics que es troben a la base del desenvolupament d'aquestes tecnologies. La manca de recursos disponibles en una llengua amb un mercat petit, com el català, encareix el desenvolupament de productes que l'integren. Aquest encariment limita l'interès que podrien tenir les grans empreses a l'hora d'incorporar el català i el fa inaccessible a les empreses emergents de l'àmbit local.

Com ja s'ha dit, en l'actual paradigma tecnològic neuronal el factor fonamental per dotar tecnològicament una llengua és l'accés a les dades i als recursos lingüístics necessaris. Això pot tenir una lectura positiva, i és que des del moment en què els recursos i les dades estan disponibles per a una llengua determinada es fa possible el desenvolupament de nous productes tecnològics per a aquesta llengua, ja siguin sistemes de traducció automàtica, sistemes de diàleg o eines de mineria de textos. La raó és que els algorismes són independents de la llengua i els avenços en llengües com l'anglès es poden transferir fàcilment a les altres llengües, si aquestes ja compten amb els recursos lingüístics necessaris.

Actualment aquests productes només existeixen per a l'anglès i unes quantes llengües més, i, per tant, només beneficien les comunitats lingüístiques que tenen el privilegi de parlar aquestes llengües. Això equival a una revolució digital a diferents velocitats segons el suport tecnològic de les llengües. Si les tendències no canvien, un gran nombre de parlants de llengües més petites no es podran beneficiar plenament de les noves tecnologies, i el futur d'aquestes llengües podria quedar greument compromès.

La dificultat d'accés a les dades és una debilitat generalitzada del sector, agreujada per la manca d'unes normes específiques d'interoperabilitat de les dades, així com de cultura de reutilització de la informació del sector públic (directiva europea RISP) en els diferents col·lectius. Calen, doncs, estratègies per garantir els recursos lingüístics de totes llengües.

9 Polítiques a nivell europeu i estatal i apoderament de la mateixa comunitat

A la Unió Europea hi ha 24 llengües oficials i el triple de llengües no oficials. En aquesta Europa multilingüe de l'era digital, la comunicació entre les persones, i entre les persones i l'Administració, així com l'accés il·limitat a la informació i al coneixement, hauria de ser possible per a tots els ciutadans europeus, independentment de la seva llengua materna, en igualtat de condicions. En el sector de les tecnologies de la llengua, els gegants americans dominen el mercat, però Europa no pot deixar la iniciativa a mans de les grans corporacions americanes, i no només per una qüestió de drets de propietat intel·lectual o confidencialitat (les dades traduïdes amb Google Translate es queden a Google), que també, sinó perquè necessita solucions específiques per als seus problemes. La investigació europea en aquest àmbit ja ha aconseguit una sèrie d'èxits notables, com ara el projecte EuroMatrix[17] o el programari de traducció automàtica de codi obert Moses; tots dos de finançament europeu. Tanmateix, la recerca finançada europea no ha aconseguit estimular prou la indústria per invertir en aquestes tecnologies. Investigadors molt qualificats, formats a universitats europees, han estat atrets per empreses no europees, fonamentalment americanes, sovint per la diferència del nivell de remuneració. En un intent de recuperar la iniciativa, el 2017 es va presentar al Parlament Europeu l'informe "Language equality in the digital age"[18] (Igualtat de les llengües en l'era digital), on es revisava la situació de desigualtat tecnològica entre les llengües a Europa i es reflexionava sobre l'obstacle que això

17 http://www.euromatrixplus.net/ [Consulta: 28 juliol 2018]

18 http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2017)598621 [Consulta: 28 juliol 2018]

suposa per al mercat digital únic, i com atempta a la igualtat entre els ciutadans. L'informe plantejava un seguit de recomanacions polítiques en cinc àrees d'actuació: institucions, investigació, indústria, mercat i serveis públics (Melero, 2018).

Val a dir que alguns països ja han iniciat les seves pròpies actuacions a nivell nacional. En concret, Espanya va aprovar el 2015 un Pla nacional per a l'avanç de les tecnologies de la llengua, amb un pressupost de 90 milions d'euros en cinc anys, amb l'objectiu de promoure la integració d'aquestes tecnologies en els processos de l'Administració pública. Aquest pla comporta l'impuls a la internacionalització d'empreses i institucions, la contractació pública de tecnologia, el suport a la creació d'infraestructures lingüístiques i la normalització i distribució dels recursos lingüístics creats per les administracions públiques. Ja hi ha contractes públics en vigor per a la utilització d'aquestes tecnologies en les àrees de justícia, turisme, salut, ciberseguretat, educació i vigilància tecnològica, i la intenció de crear un gran corpus anotat per a l'espanyol i les llengües cooficials.

La sensibilització i la implicació dels poders públics resulta fonamental en la tasca de dotar tecnològicament una llengua i garantir-ne la supervivència digital. Tanmateix, cal també la conscienciació i l'apoderament tecnològic de la mateixa comunitat de parlants, especialment si la llengua en qüestió no té estatus de llengua oficial a nivell d'estat. Els mateixos factors que posen en risc aquestes llengües els ofereixen les eines que poden fer-les sobreviure. La web s'ha anat fent cada cop més participativa, de manera que ara una gran quantitat de continguts són creats pels mateixos usuaris als blogs, els wikis i les xarxes socials. Gareth Morlais, especialista en mitjans digitals en gal·lès del govern de Gal·les, destaca en concret el paper potencial de la Viquipèdia com a mitjà per elevar el perfil tecnològic de les llengües petites.[19] Morlais afirma que grans empreses tecnològiques com Google classifiquen les llengües per la seva presència a Internet, més que pels milions de parlants que tenen. Així, per exemple, el català, que ocupa la posició 91 a la llista de llengües del món per nombre de parlants (9 milions), ocupa en canvi la posició 17 pel nombre d'articles a la Viquipèdia. Un altre exemple notable d'apoderament de la comunitat de parlants, en el cas català, és Softcatalà,[20] una associació sense ànim de lucre que té com a objectiu desenvolupar, traduir i distribuir programari en català. Softcatalà està formada per enginyers, lingüistes, dissenyadors i traductors que fan aquesta feina d'una manera desinteressada. Un altre projecte col·laboratiu de creació de recursos molt interessant és el projecte Common Voice de Mozilla,[21] que recull gravacions sonores de voluntaris amb la intenció de generar un corpus d'àudio per entrenar aplicacions que utilitzin la veu.

En conclusió, i molt especialment en el cas de les llengües amb un mercat petit, cal aplicar estratègies de suport que redueixin la inversió inicial que haurien de fer les empreses en la creació dels recursos lingüístics necessaris, i posar a l'abast de la indústria del llenguatge una infraestructura lingüística, creada amb suport públic, de dades processables, descarregables i amb una llicència que en permeti la còpia, la transformació i la creació d'aplicacions derivades. Considerem que la reutilització de dades públiques és una forma efectiva de contribuir a la creació d'aquesta infraestructura.

10 Bibliografia

Asaro, Peter (2016). The Liability Problem for Autonomous Artificial Agents. *AAAI Symposium on Ethical and Moral Considerations in Non-Human Agents* (190-194). Stanford: Stanford University.

Bel, Núria; Marimon, Montserrat (2016). Les indústries de la llengua i la tecnologia per al català. Llengua i Ús: *Revista Tècnica de Política Lingüística*, (58), 17-26.

Bel, Núria (ed.); Rigau, German (ed.) (2015). *Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España*. Madrid: SETSI.

19 https://media.ed.ac.uk/media/Welsh-language+technology+and+digital+media+-+Gareth+Morlais+at+Celtic+Knot+2017/1_snvn50d4 [Consulta: 28 juliol 2018]

20 https://www.softcatala.org/ [Consulta: 28 juliol 2018]

21 https://voice.mozilla.org/ca [Consulta: 28 juliol 2018]

Carew, Margaret; Green, Jennifer; Kral, Inge; Nordlinger, Rachel i Singer, Ruth (2015). Getting in Touch: Language and digital inclusion in Australian Indigenous communities. *Journal of Language Documentation & Conservation*, (9), 307–323.

Kornai, András (2013). Digital Language Death. *PLoS ONE*, 8(10). DOI: 10.1371/journal.pone.0077056.

Krizhevsky, Alex; Sutskever, Ilya i Hinton, Geoffrey E. (2012). ImageNet classification with deep convolutional neural networks. *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems* (1097-1105). Lake Tahoe: ACM.

Kyunghyun, Cho; Van Merriënboer, Bart; Bahdanau, Dzmitry i Bengio, Yoshua (2014). *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. arXiv e-prints, abs/1409.1259

Linguapax Review (2016). *Digital Media and Language Revitalisation. Els mitjans digitals i la revitalitzacio lingüística*. Linguapax International.

Melero, Maite (2018, 15 de novembre). Europa: la igualtat lingüística a l'era digital [entrada blog]. Consultat a https://eapc-rld.blog.gencat.cat/2018/11/15/europa-la-igualtat-linguistica-a-lera-digital-maite-melero/

Moreno, Asunción; Bel, Núria; Revilla, Eva; Garcia, Emilia i Vallverdú, Sisco (2012). *The Catalan Language in the Digital Age = La llengua catalana a l'era digital*. Heidelberg: Springer.

Moseley, Christopher (ed.) (2010). *Atlas of the World's Languages in Danger. Memory of Peoples (3rd ed.)*. París: UNESCO Publishing.

Padró, Lluís; Stanilovsky, Evgeny (2012). FreeLing 3.0: Towards Wider Multilinguality. A: *Proceedings of LREC-2012*. Consultat el 28 de juliol de 2018 a http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf

Rehm, Georg (ed.) i Uszkoreit, Hans (ed.) (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Heidelberg, Nova York, Dordrecht i Londres: Springer. Consultat a http://www.meta-net.eu/whitepapers/

Rehm, Georg; Uszkoreit, Hans; Dagan, Ido; Goetcherian, Vartkes; Ugur Dogan, Mehmet; Mermer, Coskun; Váradi, Tamás;... Gramstad, Sigve (2014). An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*. Reykjavik.

Rivera Rafael; Tarín, Carlota; Villar, Juan Pablo; Badia, Toni i Melero, Maite (2017). *Language equality in the digital age - Towards a Human Language Project*. Parlament Europeu. Colsutat el 28 de juliol de 2018 a http://www.europarl.europa.eu/stoa/en/document/EPRS_STU(2017)598621

Soria, Claudia (2017). What is Digital Language Diversity and why should we care? Dins Josep Cru (ed.), *Linguapax Review 2016: Digital Media and Language Revitalisation* (13-28). Linguapax International.

Sutherland, William (2003). Parallel extinction risk and global distribution of languages and species. *Nature*, (423), 276-279. DOI: 10.1038/nature01607