

UN RISC PER A LES NOSTRES LLENGÜES I UN REpte PER A LES TECNOLOGIES DEL LLENGUATGE

Som testimonis d'una revolució digital que té un impacte espectacular en la comunicació i la societat. Els desenvolupaments recents en la tecnologia de comunicació digital i de xarxes es poden comparar a la invenció de la impremta de Gutenberg.

La revolució digital és comparable a la invenció de la impremta de Gutenberg.

Què ens pot dir aquesta analogia sobre el futur de la societat de la informació europea i les nostres llengües en particular?

Després de la invenció de la impremta de Gutenberg, es van dur a terme grans avenços en la comunicació i en l'intercanvi de coneixement.

En els segles posteriors, s'han desenvolupat tècniques culturals per tractar millor el processament del llenguatge i l'intercanvi de coneixement:

- l'estandardització ortogràfica i gramatical de les principals llengües van permetre la ràpida difusió de noves idees científiques i intel·lectuals;
- el desenvolupament de les llengües oficials va fer possible que els ciutadans es comunicessin dins determinades fronteres (sovint polítiques);
- l'ensenyament i la traducció de llengües va permetre l'intercanvi lingüístic;

- la creació de pautes bibliogràfiques i periodístiques va assegurar la qualitat i la disponibilitat del material imprès;
- la creació de diferents mitjans com els diaris, la ràdio, la televisió, els llibres i altres formats va satisfer les diferents necessitats de comunicació.

En els últims vint anys, la tecnologia de la informació ha ajudat a automatitzar i facilitar molts dels processos:

- el programari d'edició de textos substitueix la mecanografia i la composició tipogràfica;
- El *PowerPoint* de Microsoft substitueix les transparències per retroprojector;
- el correu electrònic envia i rep documents molt més de pressa que el fax;
- L'Skype permet fer trucades telefòniques a través d'Internet i organitzar reunions virtuals;
- els formats de codificació d'àudio i de vídeo faciliten l'intercanvi de contingut multimèdia;
- els motors de cerca proporcionen accés a pàgines web basat en paraules clau;
- els serveis en xarxa com el Google Translate produeixen traduccions ràpides i aproximades;
- les plataformes dels mitjans de comunicació socials faciliten la col·laboració i permeten compartir informació.

Tot i que aquestes eines i aplicacions són útils, actualment no permeten implementar de manera suficient una societat de la informació europea sostenible i multilingüe, una societat moderna i inclusiva, on la informació i els productes puguin circular lliurement.

2.1 LES FRONTERES LINGÜÍSTIQUES DIFICULTEN LA SOCIETAT DE LA INFORMACIÓ EUROPEA

No podem saber exactament com serà la societat de la informació del futur. Quan es tracta de discutir una estratègia energètica europea o una política d'afers estrangers comunes, voldríem poder escoltar com parlen els ministres d'afers estrangers en la seva llengua materna. Voldríem poder tenir una plataforma on la gent, que parla moltes llengües diferents i amb dominis molts variats d'aquestes llengües, poguessin discutir un tema en particular mentre la tecnologia recopila automàticament les seves opinions i genera breus resums. També voldríem poder parlar amb el departament de suport o informació d'una companyia d'assegurances de salut que es troba en un país estranger.

L'espai d'economia i informació ens enfrenta amb diferents idiomes, parlants i contingut.

És clar que les necessitats de comunicació tenen una qualitat diferent en comparació a fa uns anys. Una economia global i l'espai d'informació ens confronten amb més llengües, parlants i continguts, i ens demanen una interacció més ràpida amb nous tipus de mitjans de comunicació. La popularitat actual dels mitjans de comunicació socials (*Viquipèdia*, *Facebook*, *Twitter* i *YouTube*) és només la punta de l'iceberg.

Avui en dia, podem transmetre gigabytes de text arreu del món en pocs segons abans de reconèixer que el text és en una llengua que no entenem. D'acord amb un informe recent demanat per la Comissió Europea, el 57% dels usuaris d'Internet a Europa compren productes i serveis en llengües que no són la seva llengua materna. (L'anglès és la llengua estrangera més comuna seguida del francès, l'alemany i l'espanyol.) El 55% dels usuaris llegeix continguts en una llengua estrangera mentre que només un 35% utilitza una altra llengua per escriure correus electrònics o publicar comentaris a la web [3]. Fa uns anys, l'anglès podria haver estat la lingua franca de la web —la gran majoria de continguts era en anglès— però ara la situació ha canviat radicalment. La quantitat de continguts en altres llengües (particularment en àrab i en llengües asiàtiques) s'ha disparat.

Sorprenentment, una bretxa digital omnipresent causada per les fronteres lingüístiques no ha aconseguit ser un punt de gaire interès en el discurs públic; no obstant, hi ha una pregunta a l'aire que es planteja de manera insistent: «Quines llengües europees prosperaran i persistiran en la informació en xarxa i la societat del coneixement?»

2.2 LES NOSTRES LLENGÜES EN RISC

La impremta va contribuir a un inestimable intercanvi d'informació a Europa, però també va portar l'extinció de moltes llengües europees. Poques vegades s'imprimia res en llengües regionals i minoritàries. Com a conseqüència, moltes llengües com el còrnic o el dàlmata es veien restringides a formes orals de transmissió, la qual cosa limitava la seva adopció continuada, l'extensió i l'ús.

La varietat de llengües a Europa és un dels seus béns culturals més rics i importants.

Les aproximadament 80 llengües que hi ha a Europa constitueixen un dels seus valors culturals més rics i importants. La multitud de llengües europees és també una part vital del seu èxit social [4]. Mentre les llengües populars com l'anglès i l'espanyol mantindran sens dubte la seva presència en el mercat i la societat digitals emergents, moltes llengües europees podrien quedar fora de les comunicacions digitals i esdevenir llengües irrelevantes per a la societat d'Internet; un fet que, sens dubte, no seria convenient. D'una banda, es perdria una oportunitat estratègica i com a conseqüència, la posició global d'Europa es veuria debilitada. De l'altra, s'entraria en conflicte amb l'objectiu de la igualtat de participació per a tots els ciutadans europeus, independent de la llengua. D'acord amb un informe de la UNESCO sobre el multilingüisme, les llengües són un mitjà essencial per al gaudi dels drets fonamentals, com l'expressió política, l'educació i la participació en la societat [5].

2.3 LES TECNOLOGIES DEL LLENGUATGE SÓN UNA TECNOLOGIA CLAU

En el passat, els esforços d'inversió s'han centrat en l'ensenyament de llengües i la traducció. D'acord amb algunes estimacions, per exemple, el mercat europeu de traducció, interpretació, localització de programari i globalització de llocs web era de 8.400 milions d'euros el 2008 amb un creixement anual previst del 10% [5]. No obstant, aquesta capacitat existent no és suficient per satisfer les necessitats actuals i futures.

Les tecnologies del llenguatge són una tecnologia clau que pot protegir i fomentar les llengües europees. Les tecnologies del llenguatge ajuden la gent a col·laborar, fer negocis, compartir coneixement i participar en debats socials i polítics independentment de les barreres lingüístiques o dels coneixements d'informàtica. Les tecnologies del llenguatge s'utilitzen com a ajuda per a les

tasques del dia a dia, com ara escriure correus electrònics, fer una cerca en xarxa o reservar un bitllet d'avió. Ens beneficiem de les tecnologies del llenguatge quan:

- trobem informació a través d'un motor de cerca a Internet;
- comprovem l'ortografia i la gramàtica en un processador de textos;
- mirem les recomanacions de productes en una botiga en xarxa;
- escoltem les instruccions verbals d'un sistema de navegació;
- traduïm pàgines web mitjançant un servei en xarxa.

Les tecnologies del llenguatge que es detallen en aquest article són una part essencial de les aplicacions innovadores del futur. Les tecnologies del llenguatge són típicament una tecnologia clau amb una gran ventall d'aplicacions com ara un sistema de navegació o un motor de cerca. Aquests llibres blancs se centren en la disponibilitat de tecnologies bàsiques per a cada llengua europea.

Europa necessita unes tecnologies del llenguatge robustes i assequibles per tots els idiomes europeus.

Per mantenir una posició capdavantera en la innovació global, Europa necessitarà que les tecnologies del llenguatge per a totes les llengües europees estiguin disponibles, que siguin assequibles i que estiguin perfectament integrades en entorns de programari més grans. Una experiència d'usuari interactiva, multimèdia i multilingüe no és possible sense les tecnologies del llenguatge.

2.4 OPORTUNITATS PER A LES TECNOLOGIES DEL LLENGUATGE

Les tecnologies del llenguatge poden fer que la traducció automàtica, la producció de continguts, el processament de la informació i la gestió del coneixement siguin possibles per a totes les llengües d'Europa. També poden afavorir el desenvolupament d'interfícies intuïtives per a l'ús d'electrodomèstics, maquinària, vehicles, ordinadors i robots. Tot i que ja existeixen molts prototips, les aplicacions comercials i industrials encara es troben en les primeres etapes de desenvolupament. Els èxits recents en recerca i desenvolupament han creat un ventall real d'oportunitats. Amb la traducció automàtica, per exemple, ja s'obté una qualitat molt raonable per a textos de dominis específics, i les aplicacions experimentals proporcionen informació multilingüe i gestió del coneixement, així com la producció de continguts en moltes llengües europees.

Les aplicacions lingüístiques, els sistemes de diàleg i les interfícies d'usuari basades en la veu s'han trobat fins ara en dominis altament especialitzats, i sovint ofereixen un funcionament molt limitat. Hi ha un mercat enorme d'oportunitats en l'educació i en les indústries d'entreteniment per a la integració de les tecnologies del llenguatge als jocs, a les ofertes d'entreteniment educatiu, als entorns de simulació o als programes de capacitat. Els serveis d'informació mòbils, els programaris d'aprenentatge de llengua assistits per ordinador, els entorns d'ensenyament a distància, les eines d'autoavaluació i els programaris de detecció de plagi són només uns quants exemples més dels llocs on les tecnologies del llenguatge poden jugar un paper important. La popularitat de les aplicacions dels mitjans de comunicació socials com el *Twitter* i el *Facebook* deixen entreveure que hi ha una necessitat addicional de tecnologies del llenguatge sofisticades que puguin controlar els

missatges, resumir debats, suggerir tendències d'opinió, detectar respostes emocionals, identificar les infraccions de drets d'autor o un mal ús del servei.

Les tecnologies del llenguatge representen una gran oportunitat per a la Unió Europea molt recomanable tant des del punt de vista econòmic com cultural. El multilingüisme a Europa ha esdevingut la norma. Les empreses europees, les organitzacions i les escoles també són multinacionals i diverses. Els ciutadans es volen comunicar tot traspasant les fronteres lingüístiques que encara hi ha en el Mercat Comú Europeu. Les tecnologies del llenguatge poden ajudar a superar les barres que encara queden mentre donen suport a l'ús lliure i obert de la llengua. A més a més, unes tecnologies del llenguatge multilingües i innovadores per a Europa també ens pot ajudar a comunicar-nos amb els nostres socis mundials i les seves comunitats multilingües. Les tecnologies del llenguatge donen suport a una gran quantitat d'oportunitats econòmiques internacionals.

Les tecnologies del llenguatge ajuden a superar les "traves" de la diversitat lingüística.

Un dels camps actius en recerca és la utilització de les tecnologies del llenguatge per a operacions de rescat en zones de desastres. En aquests entorns d'alt risc, la precisió en la traducció pot esdevenir una qüestió de vida o mort. El mateix raonament s'aplica a l'ús de les tecnologies del llenguatge en la indústria sanitària. Els robots intel·ligents amb capacitats lingüístiques per tractar amb diverses llengües tenen el potencial de salvar vides.

2.5 ELS REPTES DE LES TECNOLOGIES DEL LLENGUATGE

Tot i que les tecnologies del llenguatge han fet progressos considerables durant els últims anys, el ritme ac-

tual de progrés tecnològic i d'innovació de productes és massa lent.

Les tecnologies del llenguatge amb un ús molt estès, com ara les eines d'ortografia i gramàtica dels processadors de text, acostumem a ser monolingües, o només estan disponibles per a un cert grup de llengües. La traducció automàtica i els serveis en xarxa són excel·lents a l'hora de crear una bona aproximació dels continguts d'un document. Però aquests serveis en xarxa i les aplicacions professionals de traducció automàtica estan plens de dificultats quan es necessiten traduccions molt precises i completes.

A causa de la complexitat del llenguatge humà, model·lar les nostres llengües amb programaris i provar-les en el món real és un procés llarg i costós que requereix compromisos de finançament sostingut. Europa ha de mantenir el seu paper pioner enfront dels reptes tecnològics d'una comunitat multilingüe.

El progrés tecnològic necessita accelerar-se.

2.6 ADQUISICIÓ DE LA LLENGUA

Per il·lustrar com els ordinadors tracten el llenguatge i per què l'adquisició de la llengua és una tasca complicada, farem un cop d'ull a la manera com els humans adquireixen la primera i la segona llengua, i després farem un esquema de com treballen els sistemes de tecnologies del llenguatge.

Els humans adquirim les habilitats lingüístiques de dues maneres diferents. En primer lloc, un nadó aprèn una llengua escoltant la interacció entre parlants d'aquesta llengua. L'exposició a exemples lingüístics concrets per part dels seus usuaris, com els pares, els germans o altres membres de la família, ajuda els infants d'uns dos anys a

produir les seves primeres paraules o frases curtes. Això només és possible gràcies a una disposició genètica per a aprendre llengües que els humans tenen.

Aprendre una segona llengua normalment requereix molt més esforç quan el nen no es troba immers en una comunitat lingüística de parlants nadius. En edat escolar, les llengües estrangeres normalment s'adquireixen a través de l'aprenentatge de la seva estructura gramatical, vocabulari i ortografia de llibres i materials educatius que descriuen el coneixement lingüístic en termes de regles abstractes, taules i textos d'exemple. Aprendre una llengua estrangera requereix molt temps i esforç, i esdevé més difícil amb l'edat.

Els éssers humans adquirim les habilitats lingüístiques de dues maneres diferents: aprenent d'exemples i aprenent les regles subjacents de l'idioma.

Els dos tipus principals de sistemes de tecnologies del llenguatge adquireixen les capacitats lingüístiques d'una manera molt similar a la dels humans. Els mètodes estadístics permeten obtenir coneixement lingüístic a partir de grans col·leccions de textos d'exemple. Si bé és suficient l'ús de text en un sol idioma per l'entrenament, per exemple, d'un corrector ortogràfic, per l'entrenament d'un sistema de traducció automàtica han d'estar disponibles textos paral·lels en dos (o més) llengües. L'algorisme d'aprenentatge aprèn els patrons de com es tradueixen les paraules, frases curtes i oracions completes.

Aquest enfocament estadístic pot requerir milions d'oracions i augmenta el rendiment de qualitat amb la quantitat de text analitzat. Aquest és un del motius pels quals els proveïdors dels motors de cerca estan disposats a recopilar tant material escrit com sigui possible. La correcció ortogràfica en els processadors de text, la informació en xarxa disponible, i els serveis de traducció com el

Google Search i el Google Translate es basen en un enfocament (basat en dades) estadístic.

El gran avantatge dels mètodes estadístics és que la màquina aprèn ràpidament en cicles continus d'entrenament, tot i que la qualitat pot variar de manera arbitrària.

El segon enfocament de la tecnologia del llenguatge i la traducció automàtica en particular són els sistemes basats en regles. Experts en lingüística, lingüística computacional i informàtica codifiquen anàlisis gramaticals (regles de traducció) i compilen llistes de vocabulari (lexicons). Alguns dels principals sistemes de traducció automàtica basada en regles han estat objecte de constant desenvolupament durant més de vint anys. L'avantatge dels sistemes basats en regles és que els experts poden controlar més detalladament el processament del llenguatge. Això fa que sigui possible corregir sistemàticament els errors del programari i retornar informació detallada a l'usuari, especialment quan aquests sistemes basats en regles s'utilitzen per a l'aprenentatge de llengües. Però, degut a l'elevat cost d'aquests desenvolupaments, les tecnologies del llenguatge basades en regles

fins ara han estat privatives de les llengües majoritàries o d'aquelles que han tingut un recolzament institucional.

Els dos sistemes principals de tecnologies del llenguatge aprenen la llengua de manera similar.

Com que els sistemes estadístics i els basats en regles tendeixen a complementar-se, la recerca actual està treballant en sistemes híbrids que combinin els avantatges de totes dues metodologies. Aquests estudis són encara als laboratoris i han tingut poca implantació en aplicacions industrials.

En aquest capítol hem vist que moltes aplicacions ampliament usades en la societat de la informació depenen de les tecnologies del llenguatge. Degut a la seva comunitat multilingüe, això és particularment cert en l'espai econòmic i de la informació a Europa. Tot i que les tecnologies del llenguatge han progressat considerablement en els darrers anys, encara hi ha un gran potencial per a millorar la qualitat dels sistemes que treballen amb les llengües. En el que resta, descriurem l'estat actual de les tecnologies del llenguatge per al català.

The Catalan Language in the Digital Age

Rehm, G.; Uszkoreit, H. (Eds.)

2012, VI, 75 p. 24 illus. in color., Softcover

ISBN: 978-3-642-30677-8