D2.3.ca
Language Report for Catalan
(Catalan version)

Version 1.0

2011-07-27

METANET4U

www.metanet4u.eu

The central objective of the Metanet4u project is to contribute to the establishment of a pan-European digital platform that makes available language resources and services, encompassing both datasets and software tools, for speech and language processing, and supports a new generation of exchange facilities for them.
This central objective is articulated in terms of the following main goals:
Assessment: to collect, organize and disseminate information that permits an updated insight into the current status and the potential of language related activities, for each of the national and/or language communities represented in the project. This includes organizing and providing a description of: language usage and its economic dimensions; language technologies and resources, products and services; main actors in different areas, including research, industry, government and society in general; public policies and programs; prevailing standards and practices; current level of development, main drivers and roadblocks; etc.
Collection: to assemble and prepare language resources for distribution. This includes collecting languages resources; documenting these language resources; upgrading them to agreed standards and guidelines; linking and cross-lingual aligning them where appropriate.
Distribution: to distribute the assembled language resources through exchange facilities that can be used by language researchers, developers and professionals. This includes collaboration with other projects and, where useful, with other relevant multi-national forums or activities. It also includes helping to build and operate broad inter-connected repositories and exchange facilities.
Dissemination: to mobilize national and regional actors, public bodies and funding agencies by raising awareness with respect to the activities and results of the project, in particular, and of the whole area of language resources and technology, in general.

METANET4U is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META. META is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society.

METANET4U is co-funded by the participating institutions and the ICT Policy Support Programme of the European Commission

and by the participating institutions:

Faculty of Sciences, University of Lisbon

Instituto Superior Técnico

University of Manchester

University *Alexandru Ioan Cuza*

Research Institute for Artificial Intelligence, Romanian Academy

University of Malta

Technical University of Catalonia

Universitat Pompeu Fabra

Revision History

Version	Date	Author	Organisation	Description
1.0	27-07-2011	Asunción Moreno, Núria Bel, Eva Revilla, Emília García, Sisco Vallverdú	UPF and UPC	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

METANET4U

D2.3.ca
Language Report for Catalan
(Catalan version)

Document METANET4U-2011-D2.3.ca
EC CIP project #270893

Deliverable
Number: D2.3.ca

Completion: Final
Status: Submitted
Dissemination level: Public

Responsible: Asunción Moreno (WP2 coordinator)

Contributing Partners: Universitat Politècnica de Catalunya; Universitat Pompeu Fabra

Authors: Asunción Moreno, Núria Bel, Eva Revilla, Emília García, Sisco Vallverdú

Collaborative authors: Lluís Padró, José Adrián R. Fonollosa, Joan Soler, Ignasi Esquerra, Mireia Farrus

Reviewer: Paul Thompson

Índex

Resum executiu

Moltes llengües europees corren el risc de convertir-se en víctimes de l'era digital per la seva poca representació i per la falta de recursos en xarxa. Avui en dia, enormes oportunitats del mercat regional queden sense explotar a causa de les barreres lingüístiques. Si no prenem mesures ara, molts ciutadans europeus es veuran perjudicats socialment i econòmicament pel sol fet de parlar la seva llengua materna.

La tecnologia de la llengua (TL) innovadora és un intermediari que permetrà als ciutadans europeus participar en un coneixement i una societat del coneixement igualitaris, inclusius i amb una economia pròspera. La tecnologia de la llengua multilingüe serà una porta d'accés a la comunicació instantània, econòmica i fluïda i a la interacció a través de les fronteres lingüístiques.

Avui en dia, els serveis lingüístics els ofereixen principalment els proveïdors comercials dels Estats Units. El Google Translate, un servei gratuït de traducció automàtica, n'és només un exemple. L'èxit recent de Watson, un sistema informàtic d'IBM que va guanyar un episodi del concurs de televisió *Jeopardy* contra concursants humans, il·lustra l'immens potencial de la tecnologia de la llengua. Com a europeus, doncs, ens hem de qüestionar urgentment les següents preguntes:

- La infraestructura del coneixement i de les comunicacions hauria de dependre de les empreses monopolístiques?
- Podem confiar plenament en els serveis relacionats amb la llengua que poden ser desactivats per altres persones en qualsevol moment?
- Competim activament en el mercat global per a la recerca i el desenvolupament de la tecnologia de la llengua?
- Hi ha terceres parts d'altres continents disposades a tractar els nostres problemes de traducció i altres qüestions relacionades amb el multilingüisme europeu?
- Els nostres antecedents culturals europeus poden ajudar a donar forma a la societat del coneixement oferint tecnologia d'alta qualitat més bona, més segura, més precisa, més innovadora i més robusta?

Aquest llibre blanc per al català demostra que existeix una indústria de la tecnologia de la llengua i un entorn de recerca molt vius. Tot i que hi ha un cert nombre de tecnologies i de recursos per al català, el cert és que n'hi ha molts menys per al català que per a l'anglès. A més a més, la qualitat de les tecnologies existents i dels recursos és relativament pobra.

D'acord amb l'avaluació detallada que es presenta en aquest informe, és clar que s'haurien de prendre mesures immediates abans que es pugui fer un gran pas endavant en la llengua catalana.

Un risc per a les nostres llengües i un repte per a la tecnologia de la llengua

Actualment som testimonis d'una revolució digital que és comparable a la invenció de la impremta de Gutenberg.

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 vegades, 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 mitjançant esforços com el de la traducció de Luther de la Bíblia a la llengua comuna. En els segles següents, s'han desenvolupat tècniques culturals per tractar millor el processament del llenguatge i l'intercanvi de coneixement:

- l'estandardització ortogràfica i gramatical de les llengües principals 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 comuniquessin amb 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 d'escriptori 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 realitza trucades telefòniques a través d'Internet i organitza 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.

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.

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.[i] 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 dràsticament. La quantitat de continguts en altres llengües (particularment en àrab i en llengües asiàtiques) s'ha disparat.

Una economia global i l'espai de la informació ens confronten amb més llengües, parlants i continguts.

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?"

Quines llengües europees prosperaran i persistiran en la informació en xarxa i la societat del coneixement?

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.

Les aproximadament 60 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.[ii] 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 irrellevants per a la societat d'Internet; un fet que, sens dubte, no seria convenient. D'una banda, es per-

La gran varietat de llengües a Europa és un dels valors cultural més importants i una part essencial de l'èxit d'Europa.

dria 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.[iii]

La tecnologia de la llengua és 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 8.4 bilions d'euros el 2008 amb un creixement anual previst del 10%.[iv] No obstant, aquesta capacitat existent no és suficient per satisfer les necessitats actuals i futures.

La tecnologia de la llengua és una tecnologia clau que pot protegir i fomentar les llengües europees. La tecnologia de la llengua ajuda 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. La tecnologia de la llengua que s'utilitza 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 la tecnologia de la llengua quan:

La tecnologia de llengua ajuda la gent a col·laborar, fer negocis, compartir coneixement i participar en debats socials i politics a través de diferents llengües.

- 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 de la llengua que es detallen en aquest article són una part essencial de les aplicacions innovadores del futur. La tecnologia de la llengua és típicament una tecnologia clau amb una gran marc d'aplicació 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.

En un futur proper, necessitarem una tecnologia de la llengua per a totes les llengües europees que estigui disponible, que sigui assequible i que estigui perfectament integrada en entorns de programari més grans. Una experiència d'usuari interactiva, multimèdia i multilingüe no és possible sense la tecnologia de la llengua.

Oportunitats per a la tecnologia de la llengua

La tecnologia de la llengua pot 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. La tecnologia de la llengua també pot afavorir el desenvolupament d'interfícies intuïtives relacionades amb la llengua per a 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 precisió molt raonable en 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. Un dels camps actius en recerca és la utilització de la tecnologia de la llengua 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 la tecnologia de la llengua en la indústria sanitària. Els robots intel·ligents amb capacitats lingüístiques per tractar amb diverses llengües tenen el potencial de salvar vides.

Hi ha un mercat enorme d'oportunitats en l'educació i en les indústries d'entreteniment per a la integració de les tecnologies de la llengua als jocs, a les ofertes d'entreteniment educatiu, als entorns de simulació o als programes de capacitació. 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 la tecnologia de la llengua pot 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 de la llengua 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.

La tecnologia de la llengua representa 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 regla. Les empreses europees, les organitzacions i les escoles també són multinacionals i diverses. Els ciutadans es volen comunicar a través de les fronteres lingüístiques que encara hi ha en el Mercat Comú Europeu. La tecnologia de la llengua pot ajudar a superar les barreres que encara queden mentre dóna suport a l'ús lliure i obert de la llengua. A més a més, una tecnologia de la llengua multilingüe i innovadora per a Europa també ens pot ajudar a comunicar-nos amb els nostres socis mundials i les seves comunitats multilingües. Les tecnologies de la llengua donen suport a una gran quantitat d'oportunitats econòmiques internacionals.

El multilingüisme és la regla, no l'excepció.

Els reptes de la tecnologia de la llengua

Tot i que la tecnologia de la llengua ha fet progressos considerables durant els últims anys, el ritme actual de progrés tecnològic i d'innovació de productes és massa lent. No podem esperar deu o vint anys a fer millores significants que promoguin i facilitin la comunicació i la productivitat en el nostre entorn multilingüe.

El ritme actual de progrés tecnològic és massa lent per arribar a tenir productes de programari substancials en els propers deu o vint anys.

Les tecnologies de la llengua amb un ús molt estès, com ara les eines d'ortografia i gramàtica dels processadors de text, acostumem a ser monolingües, i només estan disponibles per a un cert grup de llengües. Les aplicacions per a la comunicació multilingüe requereixen un cert nivell de sofisticació. La traducció automàtica i els serveis en xarxa com el Google Translate o el Bing Translator 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. Hi ha molts

exemples coneguts de traduccions gracioses, com per exemple, les traduccions literals dels noms *Bush* (arbust, en anglès) o *Kohl* (col, en alemany), que il·lustren els reptes que la tecnologia de la llengua encara ha d'afrontar.

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 traducció automàtica—hi ha una raó per la qual el camp de la tecnologia de la llengua està estretament relacionat amb el camp de la intel·ligència artificial.

Els humans adquireixen les habilitats lingüístiques de dues maneres diferents: aprenent exemples i aprenent les regles lingüístiques subjacents.

Els humans adquireixen 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 nadons d'uns dos anys a produir les seves primeres paraules o frases curtes. Això només és possible gràcies a una disposició genètica que tenen els humans per a aprendre llengües.

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 cada vegada més difícil amb l'edat.

Els dos tipus principals de sistemes de tecnologia de la llengua adquireixen la llengua d'una manera similar a la dels humans.

Els dos tipus principals de sistemes de tecnologia de la llengua adquireixen les capacitat 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 concrets en una sola llengua o en el que s'anomena textos paral·lels, disponibles en dues o més llengües. Els algorismes d'aprenentatge automàtic modelen un cert tipus de facultat lingüística que permet obtenir patrons de com les paraules, les frases curtes i les oracions completes s'utilitzen correctament en una sola llengua o en traduir d'una llengua a l'altra. La quantitat d'oracions que es necessita per a les aproximacions estadístiques és enorme. La qualitat del funcionament augmenta a mesura que el nombre de textos analitzat també augmenta. És fins i tot habitual arribar a entrenar aquests sistemes amb textos que contenen milions d'oracions. 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.

Els sistemes basats en regles són el segon tipus principal de tecnologia de la llengua. 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. Degut a limitacions de finançament, la tecnologia de la llengua basada en regles només és viable per a llengües majoritàries.

El català a la societat de la informació europea

Aspectes generals

El català forma part de la família de llengües romàniques. La llengua catalana té uns 8 milions de parlants nadius, i hi ha gairebé 12 milions de persones que la poden parlar. És llengua cooficial en tres regions d'Espanya: Catalunya, les Illes Balears i la Regió de València, i també es parla en alguns pobles fronterers d'Aragó i Múrcia. És l'única llengua oficial d'Andorra i també es parla en el departament francès dels Pirineus Orientals (conegut com a Catalunya Nord), i a la ciutat italiana de l'Alguer a Sardenya.

L'estatus actual del català no és el mateix en tots els llocs on es parla. A Catalunya, la majoria de la gent és bilingüe. Uns estudis realitzats durant l'any 2010 per l'Institut d'Estudis Catalans confirma que el 95.3% de la població entén el català, el 60.6% el pot escriure, i el 77.5% el pot parlar, però aquesta última xifra puja al 96.4% quan l'estudi es restringeix a les persones nascudes a Catalunya. A més a més, altres estudis com el Programa per a l'Avaluació d'Estudiants Internacionals (*Programme for International Student Assessment*, PISA) 2010 revela que gairebé el 80% de la població a Catalunya pot llegir en català.

Com es dirà més endavant en aquest capítol, la llengua catalana es pot estudiar en diversos països arreu del món, especialment en universitats d'Europa i de Nord-amèrica.

Particularitats del català

La llengua catalana es distingeix clarament en dialectes, cinc dels quals (septentrional, central, nord-occidental, baleàric i valencià) tenen regles de normalització especials. Es diferencien principalment en la pronunciació d'algunes vocals, en alguns mots gramaticals (ex. articles, pronoms possessius i altres pronoms) i també en algunes paraules del lèxic.

El català utilitza vuit sons vocàlics diferents i trenta-un sons consonàntics. Utilitza l'alfabet de 26 lletres més 2 d'addicionals, la *ç* ('ce trencada') i la *l·l* ('ela geminada').

Pel que fa a l'ordre de les paraules de les oracions o expressions en català, el patró principal utilitzat és Subjecte, Verb, Objecte. No obstant, el català té una estructura molt lliure i no resulta estrany l'ús d'elements clítics que canvien l'estructura bàsica. Per exemple, l'oració: 'La Maria ens portà els regals a nosaltres' també es pot dir: 'A nosaltres ens portà els regals la Maria' o 'Els regals la Maria ens els portà'.

El català és una llengua *pro-drop*, és a dir, és possible utilitzar la conjugació del verb sense el pronom personal involucrat que fa la funció de subjecte.

En català, a diferència del francès o l'anglès, per exemple, no és possible separar construccions verbals que porten un verb auxiliar. Per exemple, en anglès podem dir: 'I had always done this'; o en francès: 'j'avais toujours fait ça'. En català diem: 'jo sempre he fet això', o 'jo he fet sempre això', però 'jo he sempre fet això' és incorrecte.

Com ja s'ha dit anteriorment, les arrels lèxiques del català provenen principalment del llatí. D'entre les llengües romàniques més

parlades, l'italià i el francès són les més properes al català, tant des del punt de vista lèxic com fonètic. Així doncs, els catalanoparlants poden entendre fàcilment l'italià o el francès.

L'ortografia en català és més transparent que en anglès, però menys que en espanyol o italià. Per exemple, vocals com la *a*, la *e* i la *o*, es poden pronunciar de manera diferent en alguns dialectes, en funció de si es troben en una síl·laba tònica o no. O la *b* i la *v*, que tenen la mateixa pronunciació en molts dialectes. Els signes d'accentuació i les dièresis s'utilitzen en català per ajudar a marcar l'accent i la pronunciació d'algunes paraules.

Desenvolupament recent

Després de la guerra civil espanyola (1936-1939) i durant la dictadura del general Franco (1936-1975) la llengua i la cultura catalanes van ser perseguides i fortament discriminades. Es va prohibir l'ús del català a l'escola, en qualsevol actuació administrativa i en qualsevol sistema de comunicació i difusió (llibres, diaris, radio, televisió i cinema).

La societat civil, no obstant, va mantenir una activitat cultural important, molts cops en clandestinitat, que va permetre un relaxament en la prohibició als anys 1970s. A partir de 1974 es comencen a publicar diaris en català i a partir de 1978 es permet l'ensenyament del català a l'escola.

"La Nova Cançó" és un moviment cultural i artístic que impulsa, a finals dels 1950s, la reivindicació de l'ús normal del català. Es forma el grup "Els setze jutges" al 1959, i els primers discs cantats en català apareixen al 1962.

Òmnium Cultural[1] va néixer al 1961 amb l'objectiu de protegir i de promocionar la cultura catalana. L'entitat es va crear en un moment històric on la cultura catalana es trobava censurada i perseguida per la dictadura franquista i per tant era una necessitat nacional recuperar-la i mantenir-la. Des d'aquest punt de vista, l'entitat va esdevenir una eina social i fonamental de resistència nacional i de suplència de les institucions catalanes, inexistents durant la dictadura. Actualment Òmnium genera debat, intervé i es posiciona davant les qüestions d'actualitat que afecten la societat civil catalana. També defensa i promou la llengua i la cultura catalanes i la identitat nacional de Catalunya.

L'any 1976 Ràdio 4, emissora pública espanyola d'àmbit regional, inicia les seves emissions exclusivament en català.

L'any 1983 TV3 fa la seva primera emissió de proves. TV3 és un canal de televisió, gestionat per l'ens públic Corporació Catalana de Mitjans Audiovisuals[2], que emet tota la seva programació en català.

La producció cinematogràfica en català es reactiva a partir de la segona dècada dels anys 1970s, i es consolida amb l'aparició de TV3, tant pel que fa a producció pròpia, com a la incorporació de subtitulat de pel·lícules i doblatge.

El 1985 la Generalitat de Catalunya i l'Institut d'Estudis Catalans funden el TERMCAT[3], un centre de terminologia de la llengua cata-

[1] http://www.omnium.cat
[2] http://www.ccma.cat

lana. La seva missió és garantir el desenvolupament i la integració de la terminologia catalana en els sectors especialitzats i en la societat en general.

Promoció de la llengua catalana

Tal i com es recull a l'article 6 de l'Estatut d'Autonomia de de Catalunya[4], la llengua pròpia de Catalunya és el Català. El català és la llengua d'ús normal en l'administració pública, mitjans de comunicació públics de Catalunya i és la llengua normalment emprada en ensenyament. El català i el castellà son llengües oficials a Catalunya. Els ciutadans de Catalunya tenen el dret i el deure de conèixer les dues llengües.

L'Institut d'Estudis Catalans[5], fundat el 1907 per Enric Prat de la Riba, té per objecte l'alta recerca científica i principalment la de tots els elements de la cultura catalana.

Amb el retorn a la normalitat a finals dels anys 70, l'Institut d'Estudis Catalans s'estructura en cinc seccions. La secció Filològica, acompleix la funció d'acadèmia de la llengua catalana que l'Institut té encomanada. Aquesta funció comporta l'estudi científic de la llengua, l'establiment de la normativa lingüística i el seguiment del procés d'aplicació d'aquesta normativa en l'àmbit que li és propi: les terres de llengua i cultura catalanes. Com a resultat de la seva funció, l'IEC publica El Diccionari de la llengua catalana, segona edició de 2007. Aquest diccionari normatiu i general és també un instrument per a conrear i fer prosperar la llengua catalana.

Enciclopèdia Catalana[6] és un projecte privat sense ànim de lucre, nascut l'any 1965, que s'ha consolidat com a referència en la publicació i consulta de en català d'obres de temàtica diversa, destacant per la seves enciclopèdies i diccionaris.

L'Institut Ramon Llull[7] es va crear l'any 2002, per part del Govern de la Generalitat de Catalunya i el Govern de les Illes Balears. Té com a finalitat la projecció exterior de la llengua catalana i de la cultura que s'hi expressa en totes les seves modalitats, matèries i mitjans d'expressió, així com la seva difusió i l'ensenyament fora del domini lingüístic tenint en compte totes les seves modalitats i variants. Disposa de dues seus, Barcelona i Palma , i d'oficines a Berlin, Londres, Nova York i París.

El Consorci per a la Normalització Lingüística[8] és un ens creat a partir de la voluntat comuna de la Generalitat i de nombrosos ajuntaments, consells comarcals i diputacions amb l'objectiu de facilitar el coneixement, l'ús i la divulgació de la llengua pròpia de Catalunya en tots els àmbits. Una de les seves funcions principals és la formació i suport lingüístic adreçat a persones de parla no catalana.

---

[3] http://www.termcat.cat/
[4] http://www.gencat.cat/generalitat/cat/estatut/
[5] http://www.iec.cat/
[6] http://www.enciclopedia.cat/
[7] http://www.llull.cat
[8] http://www.cpnl.cat/

La llengua a l'educació

A Catalunya, des de finals dels seixanta fins ara, unes 80 escoles, creades com a cooperatives de pares o professors, foren les pioneres en la recuperació de l'ús del català a l'educació. Aquestes escoles es van inspirar en la tradició pedagògica que existia abans de la Guerra Civil espanyola (1936) i van seguir el mètode de Maria Montessori.

Amb la recuperació de la democràcia després de la mort del dictador el 1975, la Constitució Espanyola del 1978 va reconèixer la pluralitat lingüística de l'estat. L'Estatut de Catalunya del 1979 i l'Estatut de les Illes Balears del 1983 van reconèixer el català com a llengua pròpia i oficial, juntament amb l'espanyol. L'Estatut de la Comunitat Valenciana del 1982 va reconèixer-ne el seu estatus com a llengua oficial amb el nom legal de valencià.

En aquells temps, després d'anys d'exclusió en l'educació, el català es va trobar en una situació de clar desavantatge respecte a l'espanyol. Per fer front a aquesta situació, els governs autonòmics van adoptar diferents estratègies.

A Catalunya, l'estratègia adoptada el 1983 fou l'anomenada immersió lingüística, inspirada en un programa dut a terme al Quebec (Canadà) per tractar amb una situació de contacte lingüístic similar al de les regions catalanoparlants d'Espanya. El model es basava en la idea que els nens no s'haurien de separar en funció de la seva llengua materna, perquè això crearia dos models d'escola diferents: un per als nens catalanoparlants i l'altre per als nens castellanoparlants. En la immersió lingüística, els nens, independentment de la llengua que parlen a casa, són escolaritzats totalment en català i aprenent a llegir i escriure en aquesta llengua. Quan els nens comencen a dominar la llengua de l'escola, s'introdueix gradualment l'espanyol en el currículum. D'aquesta manera, quan finalitza l'ensenyament obligatori, els estudiants, tenen un domini equivalent tant en català com en castellà, i són bilingües i billetrats, com revelen diversos estudis d'investigació[v].

A les Illes Balears les autoritats autonòmiques van adoptar un programa d'immersió lingüística similar al de Catalunya, mentre que el model adoptat a la Comunitat Valenciana establia diferents tipus de centres i programes en funció de la llengua materna de l'estudiant.

A França, al Departament de Pirineus Orientals (Llenguadoc-Rosselló) la situació del català és molt pitjor que la de les regions catalanoparlants d'Espanya. Tot i que el català és la llengua pròpia d'una part de la població, la llengua d'instrucció a l'escola és el francès. El 1976 es va crear La Bressola[vi] a Perpinyà, com una xarxa de vuit escoles que van adoptar el programa d'immersió lingüística en català, com a mitjà de recuperació de l'ús d'aquesta llengua en el dia a dia de la gent de la regió. A banda d'aquesta iniciativa, també hi ha algunes escoles bilingües a la regió, en les quals l'escolarització es fa en francès i en català.

L'últim informe PISA dut a terme el 2009 revelava que els estudiants a Catalunya, amb una puntuació mitjana de 498, es trobaven per sobre de la nota mitjana de l'OCDE (494) i de la nota mitjana d'Espanya (481) en relació a la comprensió lectoravii. Això significa que el fet que els nens fossin escolaritzats seguint el programa d'immersió lingüística, en el qual el català és el mitjà principal d'instrucció, no afectava el seu rendiment en la comprensió lectora de l'espanyol.

No obstant, l'informe PISA també mostra que hi ha una notable diferència entre les puntuacions obtingudes pels estudiants nadius (catalano- o castellanoparlants) i aquelles puntuacions obtingudes pels estudiants amb un rerefons migratori. Aquests resultats han reforçat la consciència pública sobre la importància de l'aprenentatge de llengües, amb una especial atenció en la integració social.

A l'inici del segle XXI hi ha un nou repte per a les escoles catalanes: l'escolarització de molts estudiants en rerefons de migració. A diferència dels anys seixanta, en què els nouvinguts eren majoritàriament castellanoparlants, actualment, els nens provenen de molts països d'arreu del món i parlen moltes llengües diferents. Per fer front a aquesta situació el govern ha creat les "aules d'acollida". Inspirades en el programa d'immersió lingüística, aquestes "aules d'acollida" es conceben com un suport temporal per als nouvinguts, mentre es troben en el procés d'adquisició de coneixements mínims per comunicar-se amb els seus companys de classe.

Aspectes internacionals

El català és una de les anomenades llengües minoritàries i ha estat reconeguda com a tal pel Consell d'Europa en el Capítol Europeu per a Llengües Regionals o Minoritàries, l'objectiu del qual és "protegir i promoure les regions històriques o les llengües minoritàries d'Europa". La importància d'aquestes llengües es testifica pel fet que hi ha més de quaranta milions de ciutadans de la UE que les parlen.

Com a llengua minoritària, el català va ser representat a l'Oficina Europea de Llengües Minoritàries, creada el 1982 per iniciativa del Parlament Europeu. L'objectiu d'aquesta organització no governamental paneuropea ha estat fomentar el respecte cap a les llengües menys protegides dins de la UE i promoure la diversitat lingüística. El català és també una de les llengües tractades a la Mercator Network[viii], una xarxa de tres centres de recerca i documentació, l'objectiu principal de la qual és esdevenir un centre de recursos especialitzat i un servei d'informació relatius a les llengües minoritzades europees. La Mercator té tres branques, cadascuna de les quals es centra en un programa temàtic: educació, legislació i mitjans de comunicació.

Si considerem totes les llengües que es parlen a Espanya, només l'espanyol té l'estatus de llengua oficial a la UE. No obstant, el novembre del 2004 el govern espanyol va lliurar a la UE la traducció de la Constitució Europea a les llengües de l'estat que també són oficials en els seus respectius territoris: català (amb el nom de català, utilitzat a Catalunya i les Illes Balears, i el nom de valencià, utilitzat a la Comunitat Valenciana), gallec i basc.

El 2005 el Consell de Ministres va reconèixer la possibilitat d'utilitzar altres llengües oficials a part de l'espanyol a les institucions europees. Després de signar alguns acords administratius amb algunes institucions de la UE, reconeixent un ús limitat i restringit del català, actualment l'estatus del català és el de llengua semioficial, una llengua de comunicació entre els ciutadans. Aquest estatus significa que els ciutadans poden escriure en català a aquestes institucions (Comissió Europea, Parlament Europeu, Consell, Defensor del Poble Europeu i Comitè de Regions), i, a la vegada, tenen el dret de rebre resposta en la mateixa llengua. Algunes publicacions i documentació oficial també es tradueixen al català. D'altra banda, a la Representació de la Comissió Europea a Barcelona, el català s'utilitza com a llengua habitual de comunicació amb els ciutadans

(campanyes d'informació, publicacions, comunicats de premsa i pàgina web).

La projecció internacional del català és força limitada. En el món de l'empresa a nivell internacional, l'ús del català és inexistent. De fet, l'anglès ha esdevingut la llengua comuna de comunicació tan a nivell escrit com oral. Actualment, des del punt de vista del client, algunes grans multinacionals utilitzen el català per tractar amb els seus clients catalans, com a valor afegit per als seus productes i com a millora dels seus serveis al client. Algunes d'aquestes empreses són Microsoft, IKEA o Toshiba.

Pel que fa a l'aprenentatge del català com a llengua estrangera, la situació és lleugerament millor. La Comissió Europea està desenvolupant un política activa en matèria de multilingüisme, l'objectiu de la qual és la preservació i la promoció de la diversitat lingüística a Europa, fomentant l'aprenentatge de llengües (incloses les llengües regionals i les minoritàries) i utilitzant el multilingüisme com a estímul per a la competitivitat. En aquest context, el Lifelong Learning Programme 2007-13 conté una selecció de projectes que promouen l'aprenentatge de llengües. Entre ells, el centre de recursos lingüístics en xarxa multilingüe Lingu@net Europa Plus[ix] dóna suport i recursos a 20 llengües europees, inclòs el català. A més a més, una decisió important presa pels representants dels Estats Membres de la UE ha estat incloure el català, així com el basc i el gallec, a la llista de llengües que s'ofereixen als cursos intensius de llengua Erasmus de l'any acadèmic 2010-2011[x]. Aquests cursos de llengua finançats per la UE tenen com a objectiu preparar els futurs estudiants d'Erasmus per al seu període d'estudi a les universitats catalanes, on s'utilitza el català com a llengua acadèmica i de comunicació.

Tot i ser una llengua minoritària, l'interès a aprendre català en universitats estrangeres es testifica pel fet que, en aquest moment (anys acadèmic 2010-11), més de 160 universitats d'arreu del món ofereixen estudis de llengua catalana[xi]. Alguns dels països en què aquesta oferta és més àmplia són Alemanya (26), Estats Units (23), Regne Unit (21), França (20) i Itàlia (17). El català es pot estudiar a 11 universitats espanyoles. L'Institut Ramon Llull (IRL)[xii], un consorci constituït pels Governs de Catalunya i de les Illes Balears, és la institució que té com a objectiu promoure la llengua i la cultura catalanes internacionalment. L'IRL és part de la Fundació Ramon Llull, creada pel Govern d'Andorra, l'IRL, el Consell General dels Pirineus Orientals, la ciutat de l'Alguer i la Xarxa de Ciutats Valencianes. Aquesta fundació involucra els governs i les institucions de l'àmbit lingüístic català, en un intent d'unificar esforços per a un objectiu comú.

L'interès del govern català i de les institucions en la projecció internacional de la llengua i la cultura també es veu reflectida en l'organització de diferents esdeveniments internacionals on el català ha estat el convidat principal o el convidat d'honor, com la Fira del Llibre de Frankfurt del 2007[xiii], el CatalanDays 2009 (Nova York)[xiv] o l'Expolangues 2010 (París)[xv]. També, en el món literari el PEN Català[xvi], fundat el 1922 (només un any després de la fundació del PEN International per C.A. Dawson Scott), ha estat una plataforma per a la projecció internacional de la literatura catalana i els escriptors de l'àmbit lingüístic català.

El català a Internet

Al contrari del que es podria esperar d'una llengua minoritària (després de tot, el català ocupa la posició 75 a la classificació de

l'Ethnologue[xvii] de les llengües segons el nombre de parlants), quan es tracta de la presència a Internet la situació és radicalment diferent. Segons Lluís Collado, la persona responsable de Google Books i Google News a Espanya i Portugal, Google situa la llengua catalana entre les 10 i 15 llengües més actives del món a la web. Google considera el català una llengua amb una activitat que va més enllà de les fronteres del seu domini lingüístic.

L'associació WICCAC[xviii], que aglutina administradors de web independents de l'àmbit lingüístic català que han creat les seves pàgines web en aquesta llengua, publica un baròmetre mensual sobre l'ús del català a Internet. El baròmetre s'actualitza mitjançant enquestes als llocs web d'empreses i organitzacions ubicades a Catalunya o altres llocs del domini lingüístic. L'enquesta també inclou les pàgines web d'empreses i organitzacions fora del domini, però que ofereixen els seus productes i serveis a Catalunya o dins del domini lingüístic. L'última actualització (abril del 2011) mostra que el percentatge global de l'ús del català és del 59.88% (mitjà). Aquest percentatge ha evolucionat lentament però amb un creixement constant des del primer baròmetre l'agost del 2002 (40.71%).

Aquesta presència del català a la web és deguda, d'una banda, a l'actitud de les institucions públiques que fomenten la normalització de l'ús d'aquesta llengua a Internet, i, d'altra banda, a les iniciatives privades de persones i organitzacions molt compromeses amb la seva llengua i la seva cultura.

Val la pena mencionar el significat i la força d'aquestes iniciatives privades, que han situat el català entre les llengües més actives a la web. Actualment, per exemple, Viquipèdia (la Wikipedia en català, amb 337.514 articles, és la 13a Wikipedia més gran en nombre d'articles. Un altre exemple és la Fundació[xix], que ha llançat una campanya a Internet per promoure la navegació en català ajudant els usuaris a configurar els seus navegadors i fer que el català sigui la llengua de navegació per defecte. En el moment de la seva creació el 2004, la Fundació puntCAT tenia com a objectiu principal la promoció de tot tipus d'activitats relacionades amb la creació, la gestió i el control del registre del domini .cat. Així, el 2005, la ICANN va aprovar la creació del domini de primer nivell patrocinat .cat, destinat a ser utilitzat per promoure la llengua i la cultura catalanes. Avui en dia, hi ha uns 50.000 llocs web que utilitzen aquest domini[xx].

El fet que Google, YouTube o Facebook, entre d'altres, ofereixin una versió en català per a les seves interfícies de navegació s'explica, almenys en part, pel pes creixen de la comunitat catalanoparlant a la web.

Una altra iniciativa molt important, pel que fa a l'ús del català a les TIC, és Softcatalà[xxi], una associació el principal objectiu de la qual és fomentar l'ús d'aquesta llengua en l'àmbit de la informàtica, a Internet i a les TIC. Softcatalà es basa en el treball voluntari d'estudiants, professionals i usuaris (informàtics, filòlegs, dissenyadors, traductors), que desenvolupen, tradueixen i distribueixen programari en català, com navegadors, eines d'Internet, programari d'ofimàtica, programari multimèdia, jocs, etc. La mitjana mensual de visitants únics ha crescut considerablement del 2006 (285.186) al 2011 (625.296), així com la mitjana mensual de visites del 2006 (689.142) al 2011 (1.366.644). El 2010 es van comptabilitzar unes 816.000 baixades de programari, entre elles les versions catalanes de l'OpenOffice (224.796) o del Mozilla Firefox (74.212).

La web també ofereix un nombre creixent de diaris digitals en català, així com alguns cursos en xarxa per aprendre la llengua.

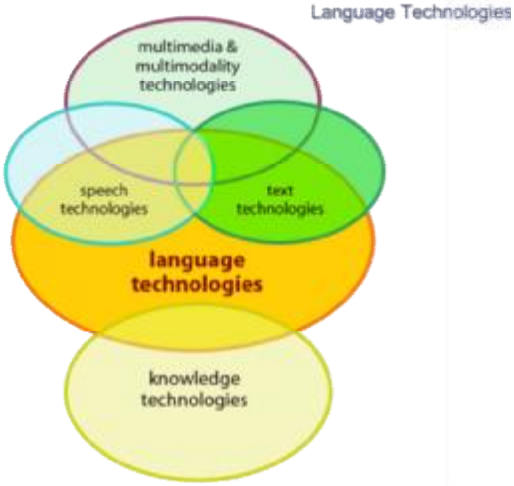
Amb tot, aquesta significant presència a Internet suggereix que hi ha una quantitat considerable de dades en llengua catalana disponibles a la web.

De totes aquestes dades, podem concloure que, tot i ser relativament petita, la comunitat catalanoparlant és molt activa i compromesa amb la seva llengua i la seva cultura, de manera que no queda exclosa de la comunicació digital. Per tant, les persones volen exercir el seu dret a utilitzar la pròpia llengua a la web en tots els nivells, ja sigui en la cerca o en la creació de continguts.

Selecció de lectures addicionals

Suport de la tecnologia de la llengua per al català

Les tecnologies de la llengua



Les tecnologies de la llengua són tecnologies de la informació especialitzades per tractar amb el llenguatge humà. Per tant, aquests tecnologies sovint s'inclouen dins del terme de Tecnologia del Llenguatge Humà. El llenguatge humà el trobem plasmat en les formes parlada i escrita. Mentre la parla és el mode més antic i més natural de comunicació lingüística, la informació complexa i la majoria del coneixement humà es mantenen i es transmeten a través de textos escrits. Les tecnologies del text i de la parla processen i produeixen la llengua en aquests dos modes de realització. Però la llengua també té aspectes compartits entre el text i la parla com els diccionaris, la major part de la gramàtica i el significat de les oracions. Així doncs, gran part de la tecnologia de la llengua no es pot incloure sota l'etiqueta de tecnologies de la parla o del text. Entre aquestes es troben totes aquelles tecnologies que enllacen la llengua amb el coneixement. La figura de la dreta il·lustra el mapa de la tecnologia de la llengua. Quan ens comuniquem barregem la llengua amb altres modes de comunicació i altres mitjans d'informació. Combinem la parla amb expressions gestuals i facials. Els textos digitals es combinen amb imatges i sons. Les pel·lícules poden contenir informació lingüística tant en la forma parlada com escrita. Així doncs, les tecnologies de la parla i del text es solapen i interactuen amb moltes altres tecnologies que faciliten el processament de la comunicació multimodal i dels documents multimèdia.

Arquitectures de les aplicacions de tecnologia de la llengua

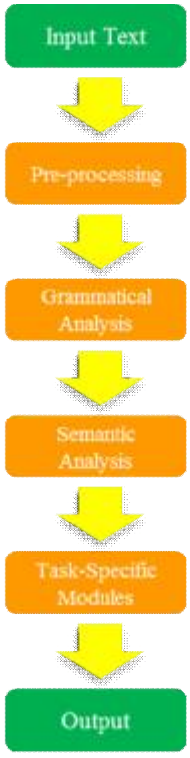


Figure 2: A Typical Text Processing Application Architecture

Les aplicacions de programari típiques per al processament de la llengua consisteixen en diversos components que reflecteixen diferents aspectes de la llengua i de la tasca que implementen. La figura de la dreta mostra una arquitectura molt simplificada tal com la podem trobar en un sistema de processament de text. Els tres primers mòduls tenen a veure amb l'estructura i el significat del text d'entrada:

- Preprocessament: netejar les dades, eliminar la formatació, detectar la llengua d'entrada, substituir "5è" per "cinquè", etc.
- Anàlisi gramatical: trobar el verb i els seus objectes, modificadors, etc.; detectar l'estructura de l'oració.
- Anàlisi semàntica: desambiguació (quin significat de *taula* és el correcte donat un context determinat?), resoldre anàfores i expressions de referència com ara *ella, el cotxe,* etc.; representar el significat de l'oració d'una manera que pugui ser llegible per la màquina.

Els mòduls específics per a cada tasca realitzen diferents operacions com ara el resum automàtic d'un text d'entrada, cerques a una base de dades, i moltes d'altres. A continuació il·lustrem les àrees principals d'aplicació i en destaquem els seus mòduls principals. Una vegada més, les arquitectures de les aplicacions són molt simplificades i idealitzades, per tal d'il·lustrar la complexitat de la tecnologia de la llengua d'una manera general i comprensible. Les eines i els recursos més importants involucrats en aquestes arquitectures es troben subratllats en el text i també es poden trobar a la taula al final del capítol. Les seccions on es discuteixen les àrees

principals d'aplicació també contenen una descripció general de les indústries actives en aquest camp a Catalunya.

Després d'introduir les àrees principals d'aplicació, mostrarem una descripció general de la situació en la recerca i l'educació sobre la tecnologia de la llengua, i conclourem amb un resum dels programes de recerca anteriors i actuals. Al final d'aquesta secció, presentarem un estimació d'experts sobre la situació per que fa a les eines i recursos de tecnologia de la llengua bàsiques en funció de dimensions com la disponibilitat, la maduresa o la qualitat. Aquesta taula mostra una bona panoràmica de la situació de la tecnologia de la llengua per al català.

Àrees principals d'aplicació

Correcció d'errors lingüístics

Qualsevol persona que hagi utilitzat algun processador de textos, com ara el Microsoft Word, haurà vist que aquests programes solen disposar d'una eina que detecta errors ortogràfics i proposa possibles correccions. Quaranta anys després del primer corrector automàtic dissenyat per Ralph Gorin, els programes actuals no es limiten a comparar les paraules utilitzades amb les que conté un diccionari, sinó que fan servir procediments cada cop més sofisticats. A més a més d'incloure algorismes dependents de la llengua que permeten analitzar aspectes morfològics (com per exemple la formació del plural), alguns poden fins i tot reconèixer errors sintàctics, com ara un verb mal conjugat (per exemple, 'Ella escrius una carta').

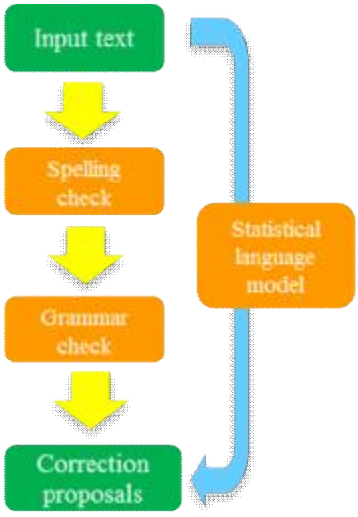


Figure 3: Language Checking (left: rule-based; right: statistical)

Per aconseguir aquestes prestacions, és necessari formular les regles gramaticals de cada llengua, una tasca manual que requereix uns grans coneixements sobre la matèria, o bé utilitzar els anomenats models estadístics de la llengua. Aquests models permeten calcular la probabilitat que una paraula concreta aparegui en un determinat entorn (és a dir, les paraules anteriors i posteriors). Per exemple, *llibre anglès* és una seqüència de paraules molt més probable que no pas *llibre anglesa*. Un model estadístic es pot obtenir automàticament a partir d'una gran quantitat de dades lingüístiques correctes (és a dir, un corpus). Fins al dia d'avui, aquests mètodes s'han desenvolupat i avaluat principalment en anglès. Malauradament, és difícil transferir-los directament al català, que té una inflexió més rica i un ordre sintàctic de les paraules més flexible.

L'ús d'eines de correcció automàtica no es limita als processadors de textos; també es poden utilitzar com a sistemes de suport per a la creació d'escrits. Degut a l'augment constant de productes tecnològics, la quantitat de documentació tècnica ha crescut ràpidament durant les últimes dècades. Per evitar possibles queixes dels clients per problemes o danys derivats d'una mala comprensió de les instruccions, les empreses han començat a posar més atenció en la qualitat de la documentació tècnica. Els avenços en el processament del llenguatge natural han permès desenvolupar programes de suport que ofereixen ajuda als autors de la documentació i els faciliten la utilització d'un vocabulari i unes estructures d'oracions consistents amb unes determinades regles i restriccions terminològiques (corporatives).

Només algunes empreses i proveïdors de serveis ofereixen productes en català en aquesta àrea. Enciclopèdia Catalana, Maxigrammar i Inèdit han creat i comercialitzat productes que inclouen la revisió ortogràfica i gramatical per al català, així com eines de revisió adaptades a diversos dominis i estils. Softcatalà i Barcelona Media

també han desenvolupat algunes eines lingüístiques que s'ofereixen a la comunitat com a aplicacions web. Una nova versió de "El corrector" s'ha desenvolupat i comercialitzat recentment per a Ipod i Ipad.

A part dels correctors ortogràfics i els programes de suport per a la creació de textos, la revisió dels diferents aspectes lingüístics també és important en l'àmbit dels programes d'ajuda per a l'aprenentatge de llengües, i s'aplica en la correcció automàtica de les cerques enviades als motors de cerca d'Internet (per exemple, en el 'Volíeu dir:' de Google).

Cerques a la web

Les cerques a la web, a intranets o a les biblioteques digitals és probablement la tecnologia de la llengua més utilitzada i, en canvi, la menys desenvolupada de totes. El motor de cerca Google, que es va posar en servei l'any 1998, s'utilitza actualment en aproximadament el 80% de totes les cerques que es fan arreu del món[xxii].

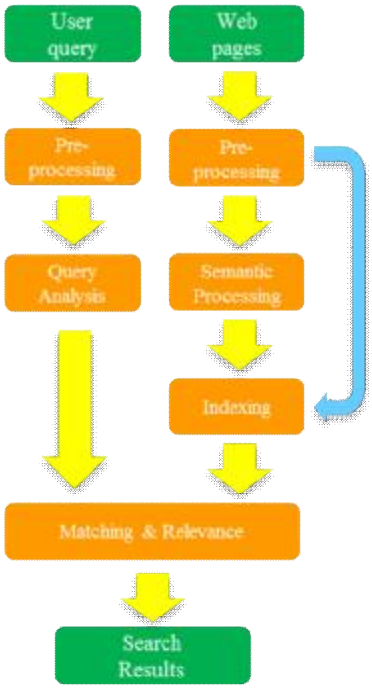


Figure 4: Web Search Architecture

Ni la interfície de la cerca ni la presentació dels resultats han canviat de forma significativa des de la primera versió. En la versió actual, Google ofereix la correcció de paraules mal escrites també en català i, des de l'any 2009, ha incorporat informació semàntica bàsica en la seva combinació algorísmica[xxiii], que permet millorar la precisió de la cerca analitzant el significat de la consulta segons els seu context. L'èxit de Google mostra que amb una gran quantitat de dades i unes tècniques eficients per indexar-les, es poden obtenir resultats satisfactoris utilitzant principalment mètodes estadístics.

Tot i això, per poder tractar cerques d'informació més sofisticades, és essencial integrar-hi un coneixement lingüístic més profund. En els laboratoris de recerca, per exemple, s'han obtingut millores en diferents experiments utilitzant tesaurus i recursos lingüístics ontològics com el WordNet que permeten trobar resultats a partir de sinònims dels termes de la consulta, o fins i tot termes més llunyans. Com passa gairebé sempre en les tecnologies de la llengua, aquests mètodes requereixen recursos específics per a cada llengua. En aquesta línia, el centre TALP de la Universitat Politècnica de Catalunya ha desenvolupat un WordNet en català, que es troba disponible de forma gratuïta[xxiv].

La propera generació de motors de cerca haurà d'incloure una tecnologia de la llengua molt més sofisticada. Si els termes d'una consulta consisteixen en una pregunta o un altre tipus de frase, en comptes d'una simple llista de paraules clau, per proporcionar respostes adequades es requereix una anàlisi sintàctica i semàntica de la consulta, a més a més de la capacitat de crear un índex que permeti trobar ràpidament els documents rellevants per a la resposta. Per exemple, en el cas que un usuari introdueixi la consulta: 'Dóna'm una llista de totes les empreses que han estat comprades per altres empreses durant els últims cinc anys'. Per proporcionar una resposta satisfactòria, és necessari aplicar una anàlisi sintàctica per tal d'analitzar l'estructura gramatical de l'oració i poder determinar que el que l'usuari desitja és saber quines empreses han estat comprades per altres, i no quines empreses han comprat altres empreses. A més a més, cal processar l'expressió 'durant els últims cinc anys' per determinar el període al qual fa referència la consulta.

Finalment, la consulta processada s'ha de comparar amb una enorme quantitat de dades no estructurades per tal de trobar la informació que l'usuari està cercant. Aquest procés es coneix habi-

tualment com a 'recuperació d'informació' i implica la cerca i la classificació dels documents rellevants. A més a més, per poder generar una llista d'empreses, cal tenir la capacitat de detectar que una cadena determinada de paraules en un document es refereix a un nom d'empresa. Aquest tipus de informació es pot obtenir amb els anomenats reconeixedors de noms d'entitats.

El fet d'intentar trobar la resposta en documents escrits en una llengua diferent de la llengua de consulta és un repte encara més gran. Per fer-lo possible, cal traduir automàticament la consulta a totes les llengües en les quals és possible trobar la resposta, i després traduir la informació trobada a la llengua original. Per altra banda, l'augment de la quantitat d'informació en formats no textuals fa que cada vegada sigui més necessària l'aparició de serveis que permetin cercar informació en entorns multimèdia, és a dir, imatges, àudio i vídeo. El cas de fitxers d'àudio i vídeo implica utilitzar sistemes de reconeixement automàtic de la parla per convertir la veu en un text en el qual es pugui cercar la informació de la consulta.

Les petites i mitjanes empreses com Inbenta (www.inbenta.com) o RightNow (anteriorment q-go, www.q-go.es) ofereixen motors de cerca amb informació semàntica en català.

També hi ha algunes iniciatives interessants per agrupar motors de cerca específics per al català, com ara http://www.cercat.cat/ o http://som-hi.com/. Es pot trobar un resum històric sobre aquest tema a: http://www.gencat.cat/diue/doc_un/cis02_partal_uk.pdf

Interacció per la parla

La tecnologia d'interacció per la parla és la base per a la creació d'interfícies que permetin als usuaris interaccionar amb màquines fent servir la veu en comptes d'un teclat, un ratolí o un altre dispositiu. Avui en dia, aquestes interfícies d'usuari orals es fan servir generalment per automatitzar, de forma total o parcial, alguns serveis que les empreses ofereixen als seus clients, treballadors i socis per via telefònica. Algunes de les àrees empresarials que depenen fortament d'aquest tipus d'interfícies són la banca, la logística, el transport públic i les telecomunicacions. Altres usos d'aquesta tecnologia consisteixen a proporcionar una alternativa a l'entrada i la sortida de dades en determinats dispositius que contenen interfícies d'usuari gràfiques, com ara els sistemes de navegació dels cotxes o els telèfons intel·ligents.

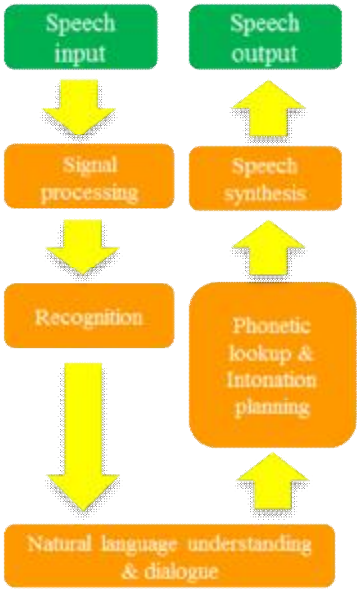


Figure 5: Simple Speech-based Dialogue Architecture

En el seu nucli, la interacció per la parla consta de les quatre tecnologies següents:

- El reconeixement automàtic de la parla és la part responsable de determinar quines paraules es corresponen amb la seqüència de sons que ha pronunciat l'usuari.
- L'anàlisi sintàctica i interpretació semàntica s'encarrega d'analitzar sintàcticament les frases que pronuncia l'usuari i d'interpretar el seu significat tenint en compte l'objectiu del sistema.
- La gestió de diàleg és necessària, en la part en què l'usuari interactua amb el sistema, per determinar quina acció s'ha de dur a terme en funció de la informació que proporciona l'usuari i de les funcions del sistema.
- La síntesi de veu s'utilitza per transformar un text en sons perquè l'usuari pugui rebre la resposta del sistema en forma de senyal d'àudio.

Un dels reptes més grans és aconseguir tenir un sistema automàtic que reconegui les paraules que pronuncia l'usuari de la manera més precisa possible. Per a això cal o bé restringir el rang de paraules possibles acceptades pel sistema a un conjunt de paraules clau, o bé crear manualment models de la llengua que donin cobertura a un ampli ventall de possibles frases que l'usuari pot pronunciar de forma natural. Mentre la primera opció ofereix una opció força rígida i inflexible que probablement es reflectirà en una baixa acceptació per part dels usuaris, la creació, l'ajust i el manteniment dels models de la llengua pot incrementar significativament els costos. Tot i això, les interfícies orals que fan servir models de la llengua i que permeten que inicialment l'usuari expressi de forma flexible la seva intenció (per exemple, saludant-lo amb un *Com puc ajudar-lo?*) mostren una taxa d'automatització més alta i una acceptació més gran per part dels usuaris, de manera que es poden considerar sistemes avantatjats respecte als mètodes, menys flexibles, de *diàleg dirigit*.

Per proporcionar una resposta als usuaris dins del marc d'una interfície oral, les empreses tendeixen molt sovint a utilitzar paraules o frases enregistrades per un locutor professional. En casos estàtics, en els quals la pronunciació no depèn del context, aquesta opció proporciona resultats satisfactoris. En canvi, si la pronunciació concreta és important per al missatge, l'experiència de l'usuari es veu molt perjudicada per la pobra prosòdia que resulta de la simple concatenació d'arxius d'àudio. Per reduir aquest problema, els sistemes de síntesi de veu actuals tenen en compte la naturalesa dinàmica de la prosòdia de les frases, i aconsegueixen uns millors, encara que millorables, resultats.

Pel que fa al mercat de la tecnologia d'interacció per la parla, l'última dècada s'ha caracteritzat per una forta estandardització de les interfícies entre els diferents components, així com dels mètodes per crear programes particulars per a determinades aplicacions. També hi ha hagut una forta consolidació del mercat en els últims deu anys, especialment en el camp del reconeixement de la parla i la síntesi de la veu. Aquí, els mercats nacionals dels països del G20 –és a dir, els països econòmicament forts i amb una població considerable– estan dominats per menys de cincs empreses, essent *Nuance* i *Loquendo* les més importants a Europa, també per al català. Tot i això, algunes empreses locals petites com Verbio[xxv], que va sorgir de la Universitat Politècnica de Catalunya i que disposa de la seva pròpia tecnologia, estan començant a competir.

Pel que fa als coneixements i la tecnologia sobre la gestió de diàlegs, els mercats estan fortament dominats per entitats nacionals, les quals són, normalment, petites i mitjanes empreses.

La majoria d'empreses en el mercat de la síntesi de veu en català són essencialment desenvolupadores d'aplicacions. Les principals dins del mercat espanyol són Indsys[xxvi] (Intelligent Dialogue Systems), Fonetic[xxvii], Ydilo[xxviii] i NaturalVoz[xxix].

Fent un cop d'ull més enllà de l'estat actual de la tecnologia, en un futur hi haurà canvis significatius degut a la proliferació de telèfons intel·ligents com a noves plataformes per gestionar les relacions amb els clients, a més a més del telèfon, Internet i el correu electrònic. Aquesta tendència també afectarà l'ús de la tecnologia d'interacció per la parla. D'una banda, a llarg termini baixarà la demanda de sistemes telefònics basats en interfícies orals. De l'altra, l'ús de la parla com a interfície amigable per als telèfons intel·ligents guanyarà importància. Aquesta suposició es veu justificada per la millora observable en la precisió dels sistemes de re-

coneixement de la parla independents del locutor que s'ofereixen actualment com a servei centralitzat per als usuaris de telèfons intel·ligents. Donada aquesta situació, l'ús d'un nucli de tecnologies lingüístiques que sigui específic per a les aplicacions guanyarà importància, suposadament, en comparació amb la situació actual.

Traducció automàtica

La idea de fer servir ordinadors per realitzar traduccions automàtiques va ser proposada per A. D. Booth l'any 1946. A partir d'aleshores, aquesta àrea va disposar de recursos substancials per a la recerca, especialment durant els anys 50 i 80. Malgrat això, la traducció automàtica encara no ha arribat a complir les altes expectatives que va generar en els seus inicis.

En la seva aproximació més bàsica, la traducció automàtica simplement substitueix cada paraula de la llengua original per la seva equivalent en la llengua objecte. Aquesta opció pot ser útil en alguns dominis amb un vocabulari molt restringit, com ara els informes meteorològics, o en casos de llengües molt properes. Tot i això, per aconseguir una bona traducció de textos menys estandarditzats, s'han d'analitzar unitats de text més llargues (frases o passatges sencers) per trobar la correspondència més adient en la llengua objecte. La dificultat més gran d'aquest aspecte és que la llengua humana és ambigua, i això implica diversos reptes, com ara trobar el sentit correcte d'una paraula ('Jaguar' pot ser un cotxe o una animal, per exemple) o la inclusió de frases preposicionals a un nivell sintàctic, com per exemple:

Passejava amb el nen cantant.

T1: I was walking with the singer child.

T2. I was walking and singing with the child.

Una manera d'afrontar aquesta tasca és basar-se en regles lingüístiques. En el cas de llengües molt properes, una traducció directa pot ser viable en casos com els de l'exemple anterior. Però sovint els sistemes basats en regles analitzen el text d'entrada i creen una representació simbòlica intermediària a partir de la qual es genera el text traduït final. L'èxit d'aquests mètodes depèn molt de la disponibilitat d'amplis lexicons amb informació morfològica, sintàctica i semàntica, i grans conjunts de regles gramàticals dissenyades amb cura per un lingüista expert.

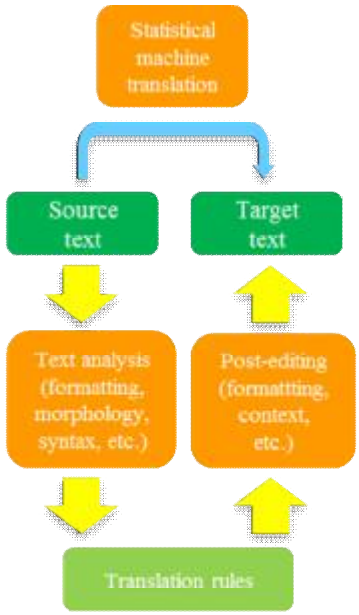


Figure 6: Machine translation (top: statistical; bottom: rule-based)

A partir de finals dels anys 80, a mesura que la capacitat computacional dels ordinadors augmentava i es feia més assequible, l'interès en els models estadístics per a la traducció automàtica va anar creixent. Els paràmetres d'aquests models es calculen a partir de l'anàlisi d'un corpus de textos bilingües, com ara el corpus paral·lel Europarl, que conté les actes del parlament europeu en onze llengües europees. Si es disposa de dades suficients, els models estadístics funcionen prou bé com per generar un text traduït que capti el significat aproximant del text original. Tot i així, a diferència dels sistemes basats ens regles, els sistemes basats en models estadístics sovint generen textos gramaticalment incorrectes. D'altra banda, a més de l'avantatge de reduir l'esforç que suposa haver d'escriure les regles gramaticals a mà, els sistemes estadístics poden cobrir particularitats de la llengua que a vegades no apareixen en els sistemes basats en regles, com per exemple les expressions idiomàtiques.

Com que els punts forts i febles d'aquests dos mètodes són complementaris, gairebé tots el investigadors treballen en sistemes

híbrids que els combinen. Això es pot fer de diferents maneres. La primera és utilitzar els dos sistemes, i després triar la millor opció per a cada frase. El problema és que per a frases llargues no hi ha cap resultat que sigui perfecte. Una solució millor és combinar les millors parts de cada frase generada amb diferents sistemes. Aquesta opció és força complexa perquè no sempre resulta obvi quines parts de cada alternativa són les que es corresponen, i per tant, s'han d'alinear.

Degut a la particular situació oficial de bilingüisme existent a les diferents regions d'Espanya i a la similitud entre el català i el castellà, els sistemes de traducció automàtica entre aquestes dues llengües funcionen de forma bastant satisfactòria. Inicialment, alguns dels principals sistemes que es van desenvolupar van ser el METAL (Siemens) i l'ATLAS (Fujitsu). Aquests projectes van tenir lloc a Barcelona durant els anys 90. Al cap d'un temps, les gran empreses que els van impulsar van apartar-se'n, i els projectes van passar a mans de diferents *spin-off*: una empresa local, INCYTA, així com GMS i Lucy Software, la qual és actualment el proveïdor principal de sistemes de traducció en català, van passar a fer-se càrrec de METAL. A més a més, el sistema de traducció entre el català i el castellà que va comprar la Generalitat de Catalunya va passar a ser un servei web públic l'any 2005, mentre que Google va començar a oferir el seu propi sistema l'any 2007.

Altres empreses, com T6 Estàndard Lingüístic i AutomaticTrans, també han desenvolupat sistemes de traducció automàtica. El sistema desenvolupat per AutomaticTrans té el seu origen en la producció d'un diari bilingüe, El Periódico. Actualment hi ha tres diaris disponibles en català i castellà que utilitzen traducció automàtica; els altres dos són El Segre i La Vanguardia.

La Generalitat Valenciana va promoure la creació del SALT, un sistema de traducció automàtica específic per al valencià. Més recentment, la Universitat Politècnica de València ha tret una versió de SiShiTra, un sistema híbrid. El sistema de codi obert OpenTrad també ofereix una versió en valencià.

La traducció entre el català i el castellà va ser l'origen de l'Apertium, un sistema de codi obert desenvolupat pel grup Transducens de la Universitat d'Alacant. L'Apertium és el primer sistema del món basat en tecnologia de traducció automàtica en codi obert, i la seva explotació comercial la duen a terme principalment Prompsit Language Engineering i OpenTrad Consortium.

La majoria d'aquests sistemes estan basats en regles. Tot i que s'estan fent esforços importants, tant a nivell nacional com internacional, per millorar els sistemes estadístics i híbrids, aquests mètodes encara no proporcionen les mateixes prestacions en aplicacions comercials que en l'àmbit de la recerca.

Donada una bona adaptació en termes de terminologia específica per l'usuari i una bona integració en el ritme de treball, l'ús de la traducció automàtica pot incrementar la productivitat significativament. Així, des de començaments del segle XXI, el principal usuari d'aquests sistemes en català és l'administració pública, inclosos alguns departaments del govern com el de justícia.

Es considera que la qualitat dels sistemes de traducció encara té un gran marge de millora. Alguns dels principals reptes són aconseguir adaptar els recursos lingüístics a un determinat domini i la integració en processos de treballs existents amb bases terminològiques i memòries de traducció.

A més a més, la majoria de corpus paral·lels existents són entre el català i el castellà, de manera que en la majoria de traductors per al català, el castellà és la llengua pivot.

La tecnologia de la llengua 'entre bastidors'

Desenvolupar aplicacions que utilitzin la tecnologia de la llengua implica una sèrie de tasques que no sempre són visibles per l'usuari final, però que proporcionen funcions importants per al sistema. Per tant, la recerca és important en aquestes àrees, les quals han esdevingut disciplines especialitzades dins de la lingüística computacional.

La cerca automàtica de respostes, tasca coneguda habitualment com a *question answering* (QA), s'ha convertit en una àrea d'investigació molt activa. Per aquest motiu, s'han construït diversos corpus degudament etiquetats i s'han posat en marxa competicions científiques. L'objectiu és passar de les consultes actuals basades en paraules clau (a les quals el motor de cerca respon amb una col·lecció de documents potencialment rellevants) a un escenari on l'usuari pugui fer una pregunta concreta i rebre del sistema una sola resposta. Per exemple, es podria preguntar: 'A quina edat va anar Neil Armstrong a la Lluna?'. I la resposta seria '38'. Tot i que aquesta tasca està evidentment relacionada amb la cerca d'informació a la web mencionada anteriorment, el terme QA s'utilitza actualment per referir-se a qüestions com el tipus de consultes que s'haurien de distingir i com s'han de tractar, com es pot analitzar un conjunt de documents que contenen potencialment la resposta cercada, i com es pot extreure la resposta d'un document de forma eficient sense deixar excessivament de banda el context.

A la vegada, tot això també està relacionat amb l'extracció d'informació, una àrea que va gaudir de molta popularitat i influència durant el principi dels anys 90, quan la lingüística computacional va començar a fer ús de mètodes estadístics. L'objectiu de l'extracció d'informació és identificar fragments específics d'informació dins de determinades classes de documents; per exemple, detectar els principals participants en compres de companyies segons les notícies aparegudes als diaris. Una altra aplicació possible és analitzar els informes sobre atemptats terroristes per extreure'n els autors, els objectius, la data i la localització de l'atemptat i els seus resultats. El fet d'haver d'analitzar textos i omplir unes plantilles que tenen un format específic del domini de la tasca concreta és una de les característiques habituals de l'extracció d'informació. Aquesta tasca és, per tant, un altre exemple de tecnologies que no són directament visibles per l'usuari perquè queden integrades en les aplicacions, però que tot i així constitueixen una àrea de recerca ben marcada.

Hi ha dues tasques, la generació i el resum de textos, que es troben al límit entre dues àrees de recerca, ja que a vegades poden ser útils directament per si soles i altres s'utilitzen com a part d'aplicacions més grans. El resum de textos es refereix, evidentment, a la tasca de crear una versió curta d'un text més llarg i s'ofereix, per exemple, com a funció dins del MS Word. Funciona principalment utilitzant mètodes estadístics, identificant primer les paraules més rellevants (que poden ser, per exemple, paraules molt freqüents en el text, però que no ho són en general) per després determinar les frases que contenen més paraules importants. Aquestes frases es marquen en el document, o s'extreuen d'ell, i s'utilitzen per construir el resum. Amb aquest mètode, que és el més utilitzat amb diferència, el resum es redueix a l'extracció d'un conjunt de frases del text principal. Tots els programes comercials que fan resums

automàtics utilitzen aquesta idea. Una alternativa, sobre la qual s'està treballant, és crear noves frases que no tenen per què aparèixer en el text original. Això requereix una comprensió més profunda del text i, per tant, és una opció molt menys robusta. En qualsevol cas, la generació de textos generalment no s'utilitza per si sola, sinó que forma part d'aplicacions més grans, com ara un sistema d'informació clínica que recull, emmagatzema i processa dades dels pacients, del qual la generació d'informes n'és només una part.

La situació actual en aquestes àrees de recerca per a la majoria de llengües es troba en un estat molt menys avançat que per a l'anglès, llengua per a la qual s'han organitzat nombroses competicions, especialment pel DARPA/NIST als Estats Units. Aquest fet ha ajudat a millorar les prestacions d'aquestes tecnologies, però principalment en anglès. També s'han fet algunes competicions en vàries llengües, però el català mai s'ha tingut gaire en compte. En conseqüència, hi ha molt pocs corpus anotats o altres recursos per a aquestes tasques. Els sistemes de resum, quan utilitzen únicament mètodes estadístics, es poden utilitzar sovint independentment de la llengua, i això ha facilitat l'existència d'alguns prototips disponibles. En el cas de la generació de text, alguns components dels sistemes en anglès es poden aprofitar, però no tots.

La tecnologia de la llengua a l'educació

La tecnologia de la llengua és un camp que involucra experts de moltes disciplines diferents com ara lingüistes, informàtics, matemàtics, filòsofs i neurocientífics, entre d'altres. Tot i que aquesta àrea de recerca encara no ocupa un lloc fix en cap universitat de l'àmbit català, és a Barcelona on es concentren la majoria de projectes.

Actualment hi ha diversos graus relacionats amb les tecnologies de la llengua disponibles en algunes universitats, com la Universitat d'Alacant, la Universitat de Barcelona, la Universitat Pompeu Fabra o la Universitat Oberta de Catalunya.

Pel que fa a màsters i postgraus, la Universitat Autònoma de Barcelona ofereix el Màster internacional en processament del llenguatge natural i tecnologies de la llengua, en col·laboració amb universitats estrangeres, i la Universitat Politècnica de Catalunya ofereix el Màster Europeu en Llenguatge i Parla. A part d'això, existeixen altres màsters en els quals el processament del llenguatge natural és una de les àrees que s'estudien (per exemple a la Universitat de Barcelona, la Universitat Pompeu Fabra, la Universitat de Girona, la Universitat Rovira i Virgili, la Universitat d'Alacant, la Universitat de Castelló i la Universitat Politècnica de València).

També hi ha programes de doctorat en els quals la tecnologia de la llengua és una de les àrees de recerca (com per exemple a la Universitat d'Alacant, a la Universitat de Barcelona i a la Universitat Rovira i Virgili).

Programes de suport per a la tecnologia de la llengua

Tant el govern català com l'espanyol han donat suport a les tecnologies de la llengua a través de diferents programes.

L'existència d'una important activitat, especialment a Barcelona, en l'àmbit de la tecnologia de la llengua es pot entendre a través dels programes de suport i projectes de recerca que s'han dut a terme en les últimes dècades. Un dels primers programes va ser

EUROTRA, un ambiciós projecte sobre traducció automàtica fundat per la Comissió Europea, que va tenir lloc entre finals dels anys 70 i l'any 1994. Encara que EUROTRA no va treballar en el català, aquest projecte, que va tenir un gran impacte en la industria de la llengua a Europa, va ser crucial per remarcar la importància de les llengües en el món de la tecnologia, i de quina manera podria afectar l'evolució tecnològica a les llengües petites i mitjanes. El govern català va entendre de seguida el repte que això implicava i va posar en marxa diversos programes per, bàsicament, localitzar diferents eines informàtiques.

La Generalitat de Catalunya ha donat suport durant més de vint anys a la recerca i el desenvolupament comercial de tecnologies de traducció automàtica, reconeixement de la parla i correcció ortogràfica i gramatical. La Secretaria de Política Lingüística, el Comissionat per a la Societat de la Informació i la Secretaria de Telecomunicacions i Societat de la Informació han estat els motors principals de les polítiques de suport. A més a més, la traducció automàtica en català s'ha beneficiat de programes de suport del govern espanyol. Projectes com l'Apertium (un sistema de traducció en codi obert) i OpenTrad, així com altres projectes petits, han rebut finançament per part del Ministeri de Ciència i Tecnologia.

El CREL (Centre de Referència en Enginyeria Lingüística, 1996-2000), gestionat per l'Institut d'Estudis Catalans amb la participació de les principals Universitats Catalanes, es va crear amb l'objectiu específic de promoure la creació d'eines i recursos pel processament automàtic de textos en català en diverses aplicacions.

Pel que fa a la presència del català en els projectes europeus, l'any 2008 el govern català va signar un acord amb la Universitat Pompeu Fabra, el representant nacional en el projecte europeu CLARIN, per construir un sistema de demostració. L'objectiu principal d'aquest sistema (CLARIN-CAT-LAB), que està disponible per a la recerca[xxx], és integrar recursos i tecnologia en català per garantir la presència d'aquesta llengua en la infraestructura resultant del projecte CLARIN. A més a més, la Biblioteca de Catalunya, juntament amb altres institucions catalanes, és un dels participants del projecte EUROPEANA.

Des de l'any 2000 fins ara, el govern espanyol ha donat suport, dins del pla nacional de recerca i tecnologia, a diversos projectes en l'àrea del reconeixement de la parla multilingüe: TEHAM, AVIVAVOZ i BUCEADOR. El principal objectiu d'aquests projectes és augmentar les prestacions dels sistemes de reconeixement de la parla, traducció i síntesi de veu en totes les llengües oficials a Espanya: basc, gallec, català i castellà.

L'any 2005, el govern català va posar en marxa un projecte per generar recursos lingüístics pel reconeixement de la parla i la síntesi de veu. Més tard, el projecte TECNOPARLA (2007-2010) es va crear per traduir veu entre el català i el castellà. Els senyals de veu es van obtenir a partir de programes de televisió. D'aquest projecte en van resultar molts avenços en diferents tecnologies: reconeixement de la parla, traducció automàtica i síntesi de veu.

Eines i recursos disponibles per al català

La següent taula proporciona un resum de la situació actual de la tecnologia de la llengua en català. La puntuació sobre els recursos i les eines existents està basat en les estimacions de diferents experts segons els següents criteris (puntuats del 0 al 6):

1. **Quantitat**: existeix un determinat recurs/eina? A més recursos/eines, més alta puntuació.
   - 0: no hi ha recursos/eines
   - 6: hi ha una gran varietat de recursos/eines
2. **Disponibilitat**: els recursos/eines són accessibles de forma gratuïta o en codi obert, o tenen un alt preu o unes condicions molt restrictives?
   - 0: pràcticament tots els recursos/eines tenen un preu molt alt
   - 6: hi ha una gran quantitat de recursos/eines disponibles de forma gratuïta sota llicències que permeten utilitzar-los lliurement
3. **Qualitat**: quina és la qualitat dels millors recursos/eines disponibles actualment? Es mantenen de forma activa?
   - 0: recursos/eines de molt baixa qualitat
   - 6: eina amb molt bones prestacions, o recurs amb anotacions equivalents a les humanes
4. **Cobertura**: fins a quin punt les millors eines cobreixen diferents criteris (estils, gèneres, longitud del text, fenòmens lingüístics, tipus de senyals d'entrada i sortida, número de llengües, etc.)? Fins a quin punt els recursos són representatius de la llengua utilitzada?
   - 0: cobertura molt baixa; només es pot utilitzar el recurs/eina en casos molt específics
   - 6: recurs amb una cobertura molt àmplia o eina molt robusta, aplicable en molts casos diversos
5. **Maduresa**: els recursos/eines es poden considerar madurs, estables i preparats per a sortir al mercat? Es poden utilitzar directament, o s'han d'adaptar primer? Les prestacions són prou bones com per utilitzar-los en aplicacions comercials o són simplement prototips per a la recerca? Un possible indicador pot ser l'acceptació del recurs/eina per a la comunitat i la seva capacitat per ser utilitzat amb èxit en sistemes reals.
   - 0: el recurs/eina és bàsicament un prototip
   - 6: component preparat per a ser utilitzat directament en aplicacions comercials
6. **Sostenibilitat**: amb quina facilitat es pot mantenir el recurs/eina i integrar-lo en els sistemes actuals? El recurs/eina compleix un determinat nivell de sostenibilitat pel que fa a la documentació, manuals, explicació d'exemples d'ús, etc.? Utilitza entorns de programació estàndards (com ara el Java EE)? Existeixen estàndards en la indústria o la recerca? En cas afirmatiu, els compleix el recurs/eina?
   - 0: sistema propietari amb formats de les dades i interfícies ad hoc
   - 6: recurs/eina completament documentat i dissenyat segons els estàndards
7. **Adaptabilitat**: amb quina facilitat es poden adaptar els recursos/eines per utilitzar-los en noves aplicacions?
   - 0: pràcticament impossible d'adaptar-los, fins i tot amb una gran quantitat d'esforç humà
   - 6: alt nivell d'adaptabilitat

Taula d'eines i recursos

	Quantitat	Disponibilitat	Qualitat	Cobertura	Maduresa	Sostenibilitat	Adaptabilitat
Tecnologia de la llengua (eines, tecnologies, aplicacions)							
Tokenització, morfologia (tokenització, desambiguació lèxica, anàlisi/generació morfològica)	4	4	5	5	5	4	4
Parsing (anàlisi sintàctica superficial o profunda)	2	3	3	3	3	3	2
Semàntica de l'oració (desambiguació del significat de la paraula, estructura argumental, rols semàntics)	1	3	2	2	2	2	2
Semàntica del text (resolució de coreferències, context, pragmàtica, inferència)	1	1	2	1	1	1	1
Processament del discurs avançat (estructura del text, coherència, estructura retòrica / teoria de l'estructura retòrica, zonificació argumentativa, argumentació, patrons de text, tipus de text, etc.)	1	1	2	2	1	1	1
Recuperació de la informació (indexació de text, RI multimèdia, RI en llengües creuades ('cross-lingual'))	3	1	3	1	3	2	2
Extracció d'informació (reconeixement de noms d'entitats, extracció d'esdeveniments/relacions, extracció d'opinions/sentiments, anàlisi/mineria de textos)	2	2	2	1	2	2	2
Generació del llenguatge (generació d'oracions, generació d'informes, generació de textos)	1	2	3	1	3	3	1
Resums, *question answering*, tecnologies avançades d'accés a la informació	0	0	0	0	0	0	0
Traducció automàtica	3	3	4	3	4	3	2
Reconeixement de la parla	3	3	3	3	3	3	2
Síntesi de veu	4	2	4	4	5	4	2
Gestió del diàleg (habilitats del diàleg i modelatge de l'usuari)	1	6	4	2	3	3	3
Recursos lingüístics (recursos, dades, bases de coneixement)							
Corpus de referència	3	3	4	3	3	3	3

	Quantitat	Disponibilitat	Qualitat	Cobertura	Maduresa	Sostenibilitat	Adaptabilitat
Corpus sintàctics (*treebanks*, arbres de dependències)	3	2	3	3	3	2	2
Corpus semàntics	2	1	1	1	1	1	1
Corpus de discurs	1	6	2	2	3	3	3
Corpus paral·lels, memòries de traducció	2	1	3	2	3	1	1
Corpus de veu (dades de veu sense tractar, dades de veu etiquetades/anotades, dades de diàleg de veu)	3	5	4	3	5	4	4
Dades multimèdia i multimodals (dades de text combinades amb àudio/vídeo)	1	4	2	2	3	3	3
Models de la llengua	1	1	3	2	4	4	3
Lexicons, terminologies	3	2	4	3	4	4	3
Gramàtiques	2	3	2	2	2	2	2
Tesaurus, WordNets	2	2	4	2	2	2	2
Recursos ontològics per al coneixement del món (ex. models superiors, dades vinculades)	2	2	3	2	2	2	2

Conclusions

La taula anterior es pot resumir amb els següents punts clau, que remarquen els aspectes més importants per poder millorar les tecnologies actuals del processament automàtic del català:

- Tot i que existeixen alguns corpus específics de gran qualitat, encara fan falta corpus grans amb anotacions sintàctiques.
- Hi ha un corpus gran en català, però accedir-hi no és fàcil ni econòmic.
- La majoria de recursos no compleixen cap estàndard. Per tant, es necessiten iniciatives per estandarditzar les dades i intercanviar formats.
- La semàntica és més difícil de processar que la sintaxi. I la semàntica dels textos llargs és més complexa d'analitzar que la de les paraules o les frases.
- Quan una determinada eina té en compte aspectes semàntics, es fa més difícil trobar les dades correctes. Per tant, es necessiten més esforços per poder dur a terme un processament profund.
- Existeixen estàndards per la semàntica a nivell de paraula (RDF, OWL, etc.); tot i així, no és senzill aplicar-los a tasques de processament del llenguatge natural.

- Actualment, el processament de la parla es troba en una fase més madura que el processament del llenguatge natural de textos escrits.
- Hi ha molts grups que treballen en la traducció automàtica, especialment entre el català i el castellà.
- S'ha aconseguit dissenyar amb èxit aplicacions particulars d'alta qualitat, però és pràcticament impossible proposar solucions estandarditzades donada la situació actual relativa al finançament.

Quant a META-NET

META-NET és una xarxa d'excel·lència creada per la Comissió Europea, actualment formada per 47 membres de 31 països europeus. META-NET agrupa l'Aliança Europea de Tecnologia Multilingüe (META en anglès), una comunitat creixent d'organitzacions i professionals de la tecnologia de la llengua a Europa.

The Multilingual Europe Technology Alliance (META)

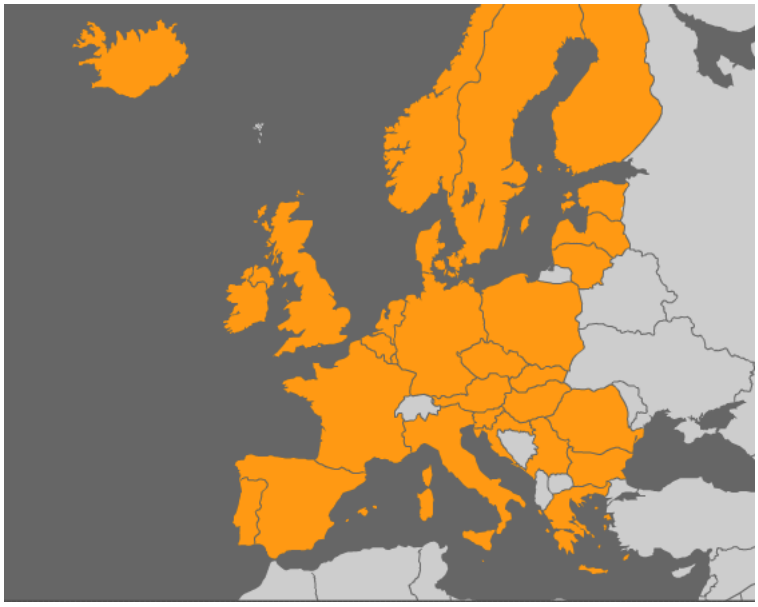


Figura 1: Països representats a META-NET

META-NET col·labora amb altres iniciatives com la Common Language Resources and Technology Infrastructure (CLARIN), i així ajuda a establir la recerca en humanitats digitals a Europa. META-NET fomenta fundacions tecnològiques per establir i mantenir una societat de la informació europea multilingüe real amb l'objectiu de:

- fer possible la comunicació i la cooperació entre diferents llengües;
- proporcionar un accés a la informació i el coneixement en igualtat de condicions per a totes les llengües;
- oferir tecnologies de la informació de forma assequible als ciutadans europeus.

META-NET estimula i promou tecnologies multilingües per a totes les llengües europees per permetre la traducció automàtica, la generació de continguts, el processament d'informació i la gestió del coneixement en una àmplia varietat d'aplicacions i dominis. La xarxa META-NET vol millorar les solucions actuals per fer més efectiva la comunicació i col·laboració en diferents llengües. Tots els europeus tenen el mateix dret a accedir a la informació i al coneixement independentment de la seva llengua.

Línies d'actuació

META-NET es va posar en marxa l'1 de febrer de l'any 2010 amb l'objectiu de fer avançar la recerca en tecnologies de la llengua. Aquesta xarxa dóna suport a la idea de crear una Europa unida pel que fa als mercats i la informació digital. A part d'això, META-NET ha dut a terme diverses activitats que van més enllà dels seus propòsits inicials. El projecte té tres línies d'acció principals anomenades META-VISION, META-SHARE i META-RESEARCH.

Figura 2: Les tres línies d'actuació de la xarxa META-NET

META-VISION agrupa una comunitat dinàmica i influent al voltant d'una visió compartida i una estratègia de recerca comuna. El principal objectiu d'aquesta activitat és construir una comunitat estable al voltant de les tecnologies de la llengua a Europa posant en contacte representants de grups diversos. Durant el primer any de META-VISION, es van fer vàries presentacions al fòrum FLaReNet (Espanya), als Language Technology Days (Luxemburg), al JIAMCATT 2010 (Luxemburg), a l'LREC 2010 (Malta), a l'EAMT 2010 (França) i a l'ICT 2010 (Bèlgica). Segons les estimacions inicials, el projecte META-NET s'ha posat en contacte amb més de 2500 professionals de les tecnologies de la llengua per tirar endavant els seus objectius. A l'esdeveniment META-FORUM, a Brussel·les, es van explicar els resultats inicials a més de 250 participants. En un conjunt de sessions interactives, els participants van compartir les seves impressions sobre les idees presentades.

META-SHARE és una xarxa oberta per intercanviar i compartir recursos. Aquesta xarxa contindrà un conjunt de dades, eines i serveis web ben documentats i organitzats en categories estandarditzades, de manera que es poden trobar i accedir-hi de forma senzilla. Els recursos disponibles inclouen tant material gratuït i de codi obert, com de pagament. META-SHARE està interessada en sistemes, dades i eines existents, així com en nous productes necessaris per construir i avaluar noves tecnologies i serveis. El fet de poder reutilitzar, combinar i modificar dades i eines per construirne de noves és un dels punts forts d'aquesta tasca. META-SHARE es convertirà en un futur en un component crucial pels professionals de les tecnologies de la llengua de petites, mitjanes i grans empreses. META-SHARE té en compte tot el cicle de desenvolupament, des de la recerca fins la creació de productes comercials. Un aspecte clau d'aquesta activitat és establir META-SHARE com una part important i valuosa d'uns infraestructura global per la comunitat de les tecnologies de la llengua.

META-RESEARCH construeix ponts entre àrees tecnològiques similars. En particular, aquesta activitat vol introduir més informació semàntica en la traducció automàtica, optimitzar el repartiment de tasques en la traducció híbrida, utilitzar al màxim la informació contextual en traduir i preparar una base empírica per a aquesta tecnologia. META-RESEARCH treballa amb altres disciplines, com

ara l'aprenentatge automàtic o la comunitat semàntica a la web. Aquesta activitat es centra en recopilar dades, construir corpus, organitzar recursos per a avaluacions, compilar eines i mètodes i organitzar congressos per a membres de la comunitat. A dia d'avui, META-RESEARCH ja ha creat recomanacions sobre com integrar informació semàntica en la traducció automàtica. A més a més, s'està acabant l'*Annotated Hybrid Sample MT Corpus*, un nou recurs que proporciona dades pels parells de llengües anglès/alemany, anglès/castellà i anglès/txec. També s'ha desenvolupat un programa que recopila dades multilingües que es troben amagades a la web.

Organitzacions membres

La taula següent mostra la llista de les organitzacions i dels seus representats que participen a la xarxa META-NET.

País	Organització	Participant(s)
Alemanya	DFKI	Hans Uszkoreit i Georg Rehm
	Universitat Tècnica d'Aquisgrà (RWTH Aachen University)	Hermann Ney
	Universitat del Saarland	Manfred Pinkal
Àustria	Universitat de Viena	Gerhard Budin
Bèlgica	Universitat d'Anvers	Walter Daelemans
	Universitat de Lovaina	Dirk van Compernolle
Bulgària	Acadèmia Búlgara de Ciències	Svetla Koeva
Croàcia	Universitat de Zagreb	Marko Tadić
Dinamarca	Universitat de Copenhaguen	Bolette Sandford Pedersen i Bente Maegaard
Eslovàquia	Acadèmia Eslovaca de Ciències Sciences	Radovan Garabik
Eslovènia	Institut Jozef Stefan	Marko Grobelnik
Espanya	Barcelona Media	Toni Badia
	Universitat Politècnica de Catalunya	Asunción Moreno
	Universitat Pompeu Fabra	Núria Bel
França	CNRS/LIMSI	Joseph Mariani
	Agència d'Avaluació i Distribució de Recursos Lingüístics (ELDA, Evaluations and Language Resources Distribution Agency)	Khalid Choukri
Grècia	Institut de la Llengua i del Processament de la Parla, "Athena" R.C.	Stelios Piperidis
Hongria	Acadèmia Hongaresa de Ciències	Tamás Váradi

País	Organització	Participant(s)
	Universitat de Tecnologia i Economia de Budapest	Géza Németh i Gábor Olaszy
Islàndia	Universitat d'Islàndia	Eirikur Rögnvaldsson
Irlanda	Dublin City University	Josef van Genabith
Itàlia	Consiglio Nazionale Ricerche, Istituto di Linguistica Computazionale "Antonio Zampolli"	Nicoletta Calzolari
	Fondazione Bruno Kessler	Bernardo Magnini
Letònia	Tilde	Andrejs Vasiljevs
	Institut de Matemàtiques i Informàtica, Universitat de Letònia	Inguna Skadina
Lituània	Institut de la Llengua Lituana	Jolanta Zabarskaitė
Luxemburg	Arax Ltd.	Vartkes Goetcherian
Malta	Universitat de Malta	Mike Rosner
Noruega	Universitat de Bergen	Koenraad De Smedt
Països Baixos	Universitat d'Utrecht	Jan Odijk
	Universitat de Groningen	Gertjan van Noord
Polònia	Acadèmia Polonesa de Ciències	Adam Przepiórkowski i Maciej Ogrodniczuk
	Universitat de Lodz	Barbara Lewandowska-Tomaszczyk i Piotr Pęzik
Portugal	Universitat de Lisboa	Antonio Branco
	Institut d'Enginyeria de Sistemes i Computadors	Isabel Trancoso
Regne Unit	Universitat de Manchester	Sophia Ananiadou
	Universitat d'Edimburg	Steve Renals
República Txeca	Universitat Carolina de Praga (Univerzita Karlova v Praze)	Jan Hajic
Romania	Acadèmia Romanesa de Ciències	Dan Tufis
	Universitat Alexandru Ioan Cuza	Dan Cristea
Sèrbia	Universitat de Belgrad	Dusko Vitas, Cvetana Krstev i Ivan Obradovic
	Institut Mihailo Pupin	Sanja Vranes

País	Organització	Participant(s)
Suècia	Universitat de Göteborg	Lars Borin
Xipre	Universitat de Xipre	Jack Burston

Referències

[i] European Commission Directorate-General Information Society and Media, *User language preferences online*, Flash Eurobarometer #313, 2011 (http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf).

[ii] European Commission, *Multilingualism: an asset for Europe and a shared commitment*, Brussels, 2008 (http://ec.europa.eu/education/languages/pdf/com/2008_0566_en.pdf).

[iii] UNESCO Director-General, *Intersectoral mid-term strategy on languages and multilingualism*, Paris, 2007 (http://unesdoc.unesco.org/images/0015/001503/150335e.pdf).

[iv] European Commission Directorate-General for Translation, *Size of the language industry in the EU*, Kingston Upon Thames, 2009 (http://ec.europa.eu/dgs/translation/publications/studies).

[v] The educational authorities in Catalonia belonged to The Network of European Language Planning Boards (http://www.languageplanning.eu/home/Pages/index.aspx)

[vi] http://www.bressola.cat/index.php

[vii] http://stats.oecd.org/PISA2009Profiles/#

[viii] http://www.mercator-central.org/

[ix] http://www.linguanet-europa.org/plus/ca/home.jsp

[x] http://ec.europa.eu/education/news/news1518_en.htm

[xi] http://www.llull.cat/_eng/_cultura/cultura_catalana_mapa.shtml?seccio=cultura&subseccio=mapa

[xii] http://www.llull.cat/_eng/_home/index.cfm?seccio=inici&ampsubseccio=1

[xiii] http://www.frankfurt2007.cat/

[xiv] http://www.llull.cat/monografics/catalandays/

[xv] http://www.llull.cat/monografics/EXPOLANGUEs/index.cfm

[xvi] http://www.pencatala.cat/

[xvii] http://www.ethnologue.com/ethno_docs/distribution.asp?by=size

[xviii] http://wiccac.cat/index.php

[xix] http://www.navegaencatala.cat/

[xx] http://www.domini.cat/

[xxi] http://www.softcatala.org/

[xxii] http://www.spiegel.de/netzwelt/web/0,1518,619398,00.html

[xxiii] http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

[xxiv] http://nlp.lsi.upc.edu/web/index.php?option=com_docman&Itemid=135

[xxv] http://www.verbio.com/

[xxvi] http://www.indisys.es/default.aspx

[xxvii] http://www.fonetic.es/

[xxviii] http://www.ydilo.com/esp/index.php

[xxix] http://www.naturalvox.com/

[xxx] http://clarin-cat-lab.org/