Estadística de la presència del català a la xarxa d'Internet i de les característiques dels llocs web catalans

Autors
Frederic Monràs, Manel Medina, Sergi Cabré, Pau del Canto, Vicenç Melendez, Enric Ripoll, Julià Urritia, Daniel Baena
Idescat i UPC

L'article exposa les característiques i avança alguns resultats de l'estudi de la presència del català a la xarxa d'Internet i de les característiques dels llocs web catalans.

Introducció

L'Idescat du a terme una estadística oficial en fase de desenvolupament per mesurar la presència de la llengua catalana a Internet i addicionalment per conèixer altres característiques dels llocs web catalans, com el tipus d'institució, els temes als quals es dediquen o els serveis que hi estan associats. L'estadística està inclosa en el Pla estadístic 2006-2009 i en el Programa anual d'actuació estadística del 2006 que concreta i desenvolupa el Pla. Es preveu incloure'l en els programes anuals següents.

El projecte es va iniciar a finals del 2003 amb la col·laboració i el suport de la Fundació Observatori per a la Societat de la informació de Catalunya (FOBSIC), que depèn de la Generalitat i la Secretaria de Política Lingüística. L'any 2006 s'hi va incorporar la Casa de les Llengües.

L'Idescat havia encarregat prèviament un dictamen de viabilitat a partir d'un avantprojecte que havia tingut com a element de referència un anterior projecte d'estadística dut a terme per la central de biblioteques OCLC l'any 1998 i següents als Estats Units. El resultat del dictamen va ser positiu i va evindeciar que el projecte era tècnicament i econòmica factible. Posava com a requisit la intervenció d'una entitat de recerca que aportés els coneixements precisos, molt especialitzats, i els instruments tècnics adequats, en forma d'equipament informàtic, de telecomunicacions i programari. Feia una especial incidència en la necessitat d'assegurar el manteniment de l'assessorament del projecte i, per tant, la seva continuïtat al llarg del temps. Per aquests motius es va encarregar el disseny i la realització al Centre d'Aplicacions d'Internet de la UPC, a Canet.

Es fan servir els sistemes d'enquesta, habituals en l'estadística oficial, basats en un mostreig aleatori fet sobre una base de dades o un cens d'individus que cobreixi exhaustivament els llocs web d'Internet expressats pels noms de domini registrats.

Amb aquest instrumental s'alleugereix l'obtenció de dades, atès que l'estudi no exigeix abastar el conjunt de llocs d'Internet, i, en canvi, es pot obtenir un resultat força ajustat, amb uns marges d'error de mostreig realment petits.

D'altra banda, es disposa dels instruments essencials per dur a terme el projecte, en concret, del programari per identificar automàticament les llengües. L'estudi d'altres dades i característiques, referents a una àrea territorial com Catalunya, és un important subproducte d'aquesta primera anàlisi lingüística a partir de la qual és possible delimitar els llocs web de més probable localització administrativa a Catalunya (la majoria dels quals en català i castellà).

La idea clau és fer una estadística amb una metodologia clara i constatable per poder disposar per primer cop d'una informació fiable. Fins ara han proliferat i s'han utilitzat moltes dades, de les quals, però, se'n desconeixen la metodologia amb què s'han obtingut, la qualitat que tenen i la data de referència.

Un altre aspecte bàsic que cal destacar és que, en el procés d'obtenció de dades sobre la llengua catalana, també es pot recollir informació sobre la presència d'altres llengües, amb la qual cosa el projecte té una transcendència no tan sols dins l'àmbit català sinó també en l'internacional.

Les llengües a Internet

Al món hi ha entre 4.500 i 6.000 llengües (en números rodons), segons què s'entengui per llengua i per variant dialectal i/o subdialectal.

Es considera igualment que només hi ha unes 800 llengües amb transcripció escrita (escriptura o *script*, en anglès) a tot el món, i d'aquestes, només unes 400 tenen codi de llengua específic (ISO 639, de tres lletres, entre d'altres).

Com a nota il·lustrativa, s'ha publicat la declaració universal dels drets de l'home en no gaire més de 350 escriptures (no podem dir-ne aquí ja llengües, atès l'exemple revelador del xinès, ja que una sola codificació escrita, tant en la forma clàssica com en la simplificada, agrupa més de catorze llengües oficials i més de cent dialectes, deixant de banda que és parlada per 1.600 milions de persones a tot el món). Així

mateix, una llengua, per exemple l'àrab, es pot escriure en un *script* aràbic o llatí (cas del maltès). Això passa també en el cas d'alguns idiomes eslaus amb els alfabets ciríl·lic i llatí, i també en altres llengües del món.

Addicionalment, aquests *scripts* es poden representar informàticament segons diferents jocs de caràcters, relacionats de vegades amb els sistemes operatius o les màquines amb què s'han generat.

Tot i que Internet és un entorn universal, en el qual totes les llengües coexisteixen virtualment una a la vora de l'altra, la manera de representar les llengües és molt diferent en diverses parts del món, segons les diverses cultures. Hi ha llengües que es representen de manera ideogràfica, com la xinesa, amb un nombre d'elements o ideogrames que pot oscil·lar entre els 2.000 (per llegir un diari) i els 12.000 (una persona culta), dins d'un horitzó molt genèric d'uns 100.000 (conjunt de tots els ideogrames existents en el món xinès). Hi ha llengües que es representen per mitjà de síl·labes (hindi) i que poden tenir entre 100 i 1.000 elements. I encara n'hi ha d'altres que són consonàntiques, basades en unes quantes desenes de combinacions (hebreu, àrab), o vocalicoconsonàntiques (les ciríl·liques i llatines, entre moltes d'altres).

La ISO 10646 i Unicode són esforços de normalització dels nombrosos jocs de caràcters, i pretenen que les llengües es representin amb aquest únic sistema de codificació, amb independència del tipus de llengua que sigui. Tenen capacitat per representar els ideogrames, que exigeixen un volum de símbols important, atès que es fa servir un format de 16 bits, la qual cosa permet aconseguir fins a 65.536 entitats diferents. I permeten també tenir en compte versions lingüístiques antigues i llengües mortes.

En un terreny més operatiu, considerant les llengües que es poden trobar a Internet, dins del nostre projecte, hem fet una agrupació en 45 superfamílies de llengües aprofitant les codificacions provinents de les normatives que s'apliquen en el món d'Internet.

Il·lustrativament, diguem que el sistema operatiu Windows XP identifica 75 idiomes, mentre que Google en té en compte 35 (entre els quals hi ha el català).

Idíl·licament, els continguts textuals que trobem a Internet segons els estàndards HTTP/1.0/1.1 —normativa que permet el diàleg navegador-servidor web— haurien de contenir en llurs capçaleres *sempre* i dins d'unes etiquetes específiques dues coses: el codi de l'idioma principal contingut en la plana web descarregada, i el codi o bé el nom del joc de caràcters amb el qual està codificada.

Aquest elements, llengua i joc de caràcters, s'hi introdueixen en forma de declaracions —seguint, en teoria, la normativa RFC 4646— en les pàgines web HTML, XHTML, DHTML, o procedents de servidors que generen planes dinàmiques (ASP, ASPX, PHP, d'altres CGI genèrics, etc.).

La realitat ens ha demostrat que estem molt lluny d'aquesta situació ideal; en la majoria de casos manquen una o ambdues codificacions, cosa que força un processament per identificar la llengua, per a la qual cosa cal fer en primer lloc una tasca de neteja amb l'extracció de les marques de codificació de plana web (etiquetes HTML) i els *scripts* de la part programàtica (JavaScripts, vbScripts, Java entre d'altres), abans de poder procedir al processament lingüístic del contingut textual restant, segons una escala de mètodes.

Cal considerar, a més, que un document pot tenir diverses parts en llengües diferents.

Les dades de llengües a Internet han sovintejat, com s'ha dit més amunt, però la seva fiabilitat i transparència metodològica no ha estat suficient. La UNESCO té un interès especial en aquestes dades, ja que han de servir per diagnosticar un previsible procés d'uniformitat cultural *de facto* conseqüència d'Internet i també per afavorir la màxima multiculturalitat i verificar-la.

Metodologia

La base de dades per a l'extracció de les mostres

Com s'ha dit més amunt, el procediment previst per l'obtenció de les estadístiques previstes en el projecte és treure una mostra aleatòria d'una base de dades representativa del contingut d'Internet i analitzar-la seguint la pautes habituals dels instituts d'estadística.

En aquest punt s'ha seguit en part el precedent, és a dir, l'experiència de l'OCLC, que va fer un mostreig a partir de números d'IP. Aquest sistema, si bé era vàlid al començament, produïa biaixos a causa de la implantació de l'estàndard HTTP/1.1 amb el qual s'aconsegueix l'adreçament a dominis virtuals; això vol dir que una adre-

ça numèrica única d'internet (d'IPv4 o bé d'IPv6), complementada amb un «nom de domini», dirigeix cap a un lloc web d'Internet amb els seus continguts separats; per tant, molts dominis poden adreçar-se a un sol servidor web (*multihoming*).

L'enfocament proposat no es basa, doncs, en els números IP com a individus a mostrejar, sinó que se substitueix pels noms de domini. Per tant, la unitat de comptatge que cal emprar és el «nom de domini», és a dir, qualsevol nom que es registra per accedir a llocs web d'Internet. Cal tenir present que un lloc web pot estar format per un o més noms de domini, sigui perquè està dividit en parts i cadascuna té el seu nom, sigui perquè hi ha diversos noms que dirigeixen cap a la mateixa informació (per exemple, idescat.net, idescat.org, etc., en el segon cas, i mediambient.gencat.net i gencat.net, en el primer).

Es va descartar inicialment produir la base de dades de referència amb noms de domini, atès que es preveia massa costós, i es va decidir d'adquirir-ne una de suposadament exhaustiva i de la qual es coneguès el procediment de producció. Es va optar per la base de dades de l'Internet Software Consortium (ISC), que a l'octubre del 2003 tenia ja més de 290 milions de noms. Es va descartar Netcraft, tot i que hi havia exemples d'utilització de la base per part de l'OCDE amb fins estadístics, a causa de la incertesa amb relació a la seva metodologia i dels costos associats.

L'opció escollida no va ser bona, ja que ens van subministrar una base de dades que era en realitat una mescla de noms de domini finals barrejats també amb noms de servidors de *middleware* de telefòniques, *routers*, noms de servidors de DHCP de telefòniques per a l'enrutatge de connexions ADSL, entre d'altres. És a dir, equipament intermedi i no final. En conseqüència per tal d'evitar una depuració molt costosa, es va optar per fer una base de dades pròpia a partir d'un procés reiterat —cíclic— de recopilació d'enllaços trobats mitjançant l'accés a un nucli inicial de llocs web —triats expressament— i amb la qual cosa s'obviaven aquest noms de domini, auxiliars, no finals, atès que no apareixien com a enllaç.

També es va descartar fer les estadístiques a partir de dades obtingudes de cercadors (Google, Alltheweb, Altavista, Vivissimo, entre d'altres). Deixant de banda les qüestions jurídiques i de possible propietat empresarial dels seus continguts, la seva finalitat no és indexar de manera representativa el contingut d'Internet: tenen un propòsit específicament comercial, a més d'estar subjectes a canvis en la política d'indexació (posicionament segons rànquings via pagament), no són exhaustius, ni controlen totes les llengües de la mateixa manera i finalment no publiquen la metodologia que empren.

Per tant, la base de dades mestra ha crescut durant el procés de retroalimentació d'ella mateixa amb nous noms de domini identificats dins del contingut dels lloc web als quals s'ha accedit.

S'han comparat els totals per TLD i per ccTLD amb altres estudis existents, com per exemple els publicats per SecuritySpace, amb resultats congruents i que globalment li donen validesa.

Durant la seva creació s'ha arribat a uns percentatges d'actualització constants i s'ha procurat que la major presència d'enllaços en algunes llengües no produïssin biaixos cap a aquestes en detriment d'altres que se citin menys. Es procedeix igualment a donar de baixa noms de domini desapareguts.

En general, el factor «gran volum de dades a manipular», bases de dades molt grans, mostres molt grans, descàrregues de pàgines molt grans, etc., han provocat canvis i millores en el projecte i al mateix temps l'han fet més robust. S'ha de tenir present que la primera mostra corresponent al 2005 amb text pla sense continguts de fitxers multimèdia[1] ocupava 2 TB (uns 2.000 GB), mentre que el volum de la descàrrega del primer semestre del 2006 ja era de 4,7 TB.

La generació de mostres es fa un parell de cops l'any. Enguany ja se n'ha fet una al juliol i se n'està preparant una altra al desembre.

Procediment d'identificació de les llengües

Evidentment, per analitzar en concret els llocs web amb presència del català més a fons i per identificar i verificar que efectivament hi ha text en català, cal un progra-

1. Es consideren fitxers d'aquest tipus els mp2, mp3, ra, PDF, PS, AVI, MPEG, DOC, XLS..., i uns quants centenars més.

mari, del qual ja es disposa, per analitzar de manera ràpida i eficient un determinat contingut. És també en aquest moment que es recolliran altres característiques del lloc web, com el tipus d'institució productora, el tema, els serveis, etc.

El nombre de pàgines que es baixen de cada nom de domini seleccionat en la mostra és al voltant de 2.000. Això obliga a plantejar-se el sistemes d'emmagatzematge, de neteja de codis en el text, de compressió i descompressió ràpides, més eficients a fi que no es generin retards. Gràcies a aquesta possibilitat de manipulació de dades es possible dur a terme les anàlisis esmentades.

Al marge de la mostra, es va obtenint igualment un cens de tots els noms de domini que contenen el català, als efectes de donar solidesa a les dades d'idiomes i de qualsevol naturalesa que es vagin obtenint de les mostres. Aquest cens es pot dur a terme perquè la quantitat de noms de domini a obtenir de la llengua catalana és assequible.

L'anàlisi de la llengua es basa inicialment en les indicacions de llengües presents en les pàgines de text pla. En cas que no n'hi hagi, es mira amb quins jocs de caràcters és compatible el que s'ha trobat en les pàgines. Igualment, per a les llengües en alfabet llatí hi ha algoritmes (n-grams, grups, normalment de tres lletres consecutives, que presenten unes pautes regulars segons quina sigui la llengua) que identifiquen amb molta precisió de quina llengua es tracta, sempre que hi hagi un volum de text suficient (a partir de 50 paraules). Això permet diferenciar llengües molt estretament emparentades, com és el cas del català amb altres llengües llatines.

Tecnologia

El sistema informàtic dissenyat per descarregar, emmagatzemar i analitzar les pàgines dels noms de domini obtinguts per mostreig, consta de dos elements clau:

1. Entorn de xarxa de computació (*grid computing*) que opera sobre un conjunt d'ordinadors PC

2. Robot per a la descàrrega de llocs web

El primer, i molt succintament, està compost d'un programa anomenat «consola de Grid», un programari de control o gestor que controla el funcionament del conjunt de màquines i l'emmagatzematge dels resultats obtinguts; en concret, s'ocupa de controlar tots els nodes que componen la xarxa de computació, el *middleware* de la xarxa (*switch*, *router, et altri*) i els nodes de computació que es corresponen a ordinadors (CPU) complets que processen tot allò encarregat i auditat per la consola.

El segon és un robot[2] d'Internet, o sigui, un programari que amb passi de comandes interactua de manera mimètica com si fos un navegador (eficaç, per tant, per descarregar llocs web compatibles i/o bé preparats per ser visualitzats per navegadors com ara l'MS Internet Explorer v4, Opera v5 i superiors, Netscape v4 i superiors i variants de Mozilla —Firefox—) que accedeix al domini i en descarrega els continguts i n'interpreta els enllaços de manera totalment automàtica (d'aquí deriva el nom, de manera robotitzada).

Segons els protocols acordats amb l'Idescat el robot descarrega un nombre de pàgines segons unes àlgebres predeterminades i que no superen les dues mil per lloc web repartides segons els diversos nivells del nom de domini (nivells de directoris, si és que hi són presents).

Es requereix que el robot funcioni bé en qualsevol lloc web i que no sigui percebut com una amenaça per al lloc accedit, de manera que no se li impedeixi el pas.

Els robots han d'actuar en paral·lel i en un nombre de màquines suficient per dur a terme l'operació de descàrrega d'una mostra en un temps raonable, que actualment és de menys d'un mes.

Cada ordinador o node del Grid pot executar diversos robots en paral·lel, i cadascuna d'aquestes execucions descarrega de manera simultània cent llocs web (a part d'empaquetar les dades en un 10 % de la seva grandària, que immediatament són transmeses als gestors documentals d'emmagatzematge massiu).

La connectivitat a Internet la proveeix el servei estàndard de la UPC.

L'anàlisi de la llengua s'efectua en una segona fase, un cop s'han descarregat els noms de domini, que se suposa que seran suficients per poder disposar d'un volum d'uns 2.000 noms que continguin text en llengua catalana.

---

2. Robots d'Internet també reben el nom de *bots*, *crawlers*, *spiders*, entre altres denominacions.

Resultats

La primera mostra del juliol del 2006 contenia un total 682.728 noms de domini amb identificació d'algun dels 142 codis de llengües buscats. Per a la comptabilització del català, castellà i anglès, s'ha tingut en compte la declaració directa de l'idioma, però s'ha fet també una recerca exhaustiva en tots els codis d'idiomes declarats susceptibles d'incloure'ls (és a dir, a tot arreu excepte en les pàgines declarades amb codis d'idioma unívocs, com el xinès, el japonès, l'àrab, etc.). En el cas de l'anglès, s'ha pres la mesura addicional d'excloure les pàgines o els noms de domini que tenen exclusivament una cinquantena de paraules en anglès d'ús corrent en qualsevol idioma (*software*, *homepage*, etc.).

En aquesta mostra, el català tindria un percentatge del 0,16 % en termes de pàgines i del 0,27 % en termes noms de domini. Els mateixos conceptes per al castellà serien 1,29 % i 1,48 %, i per a l'anglès —amb l'afegit del codi de caràcters declarat com a ASCII— les dades serien 41,0 % i 32,82 %, respectivament. Es tracta d'un avanç de les dades que se subministraran del 2006; per tant, cal tenir en compte que es tracta d'una estadística encara no consolidada.

■