

El català i les tecnologies de la llengua

Gemma Boleda⁽¹⁾, Montse Cuadros⁽¹⁾, Cristina España-Bonet⁽¹⁾,
Maite Melero⁽²⁾, Lluís Padró⁽¹⁾, Martí Quixal⁽²⁾, Carlos Rodríguez⁽²⁾

⁽¹⁾ Centre de Recerca TALP
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

⁽²⁾ Fundació Barcelona Media - Universitat Pompeu Fabra

{gboleda,cuadros,cristinae,padro}@lsi.upc.edu
{maite.melero,marti.quixal,carlos.rodriguez}@barcelonamedia.org

RESUM

El processament computacional de la llengua abraça qualsevol activitat relacionada amb la creació, gestió i utilització de tecnologia i recursos lingüístics. En el pla científic, aquesta activitat és central en disciplines com ara la lingüística de corpus, l'enginyeria lingüística, o el processament del llenguatge natural escrit o parlat. En el pla quotidià, s'inclou en un ampli ventall d'aplicacions cada cop més habituals: sistemes automàtics d'atenció telefònica, traducció automàtica, etc.

La gran majoria d'aquestes aplicacions requereixen eines i recursos lingüístics específics per a cada llengua. Per a llengües amb un mercat ampli, com l'anglès o el castellà, l'oferta de productes i serveis basats en tecnologia lingüística és variada i habitual. Per al cas de llengües com el català, és més difícil trobar productes i serveis que s'ofereixin ja “de fàbrica” amb aquesta tecnologia.

Aquest article presenta una panoràmica de l'estat actual de les tecnologies de la llengua per al català, així com diversos aspectes que avui dia es debaten en el si de la comunitat científica dedicada al processament del llenguatge natural parlat i escrit.

1. INTRODUCCIÓ

Cada cop són més habituals els productes tecnològics que incorporen un cert grau de processament computacional de la llengua. Des del senzill predictor de paraules del telèfon mòbil, fins a un traductor automàtic integrat al navegador o al correu electrònic, les interfícies tecnològiques intenten apropar-se cada cop més a la forma d'interacció més natural per a les persones: la llengua.

El desenvolupament d'aquestes tecnologies és sovint costós, especialment pel que fa a les eines i recursos que són específics per a cada llengua. Per tant, cada llengua que s'afegeix als productes suposa una inversió que no sempre resulta rendible per a les empreses que els desenvolupen. En el cas de llengües com el català, és clau el paper de les administracions, universitats, i centres de recerca a l'hora de facilitar l'accés a aquestes eines i recursos per tal que desenvolupar aquests productes en català suposi una rendibilitat comparable a fer-ho en una altra llengua amb un mercat més ampli.

La visió de l'estat actual de les tecnologies de la llengua per al català i de les qüestions controvertides relacionats que presentem en aquest article emanen dels continguts de la primera Jornada de Processament Computacional del Català (JPCC)¹, que va tenir lloc el març del 2009 amb els objectius de (1) millorar la comunicació i la col·laboració entre els diferents grups de recerca,

¹ Remetem el lector al lloc web de la JPCC per accedir a tota la documentació generada al voltant de la Jornada i que pot complementar la informació presentada en aquest article: <http://sites.google.com/site/jornadacatala/>.

empreses i institucions que desenvolupen eines i recursos computacionals per al català, (2) trobar maneres d'aprofitar de forma eficient els recursos existents i, (3) donar visibilitat a la recerca en el tractament computacional del català.

La secció 2 d'aquest article presenta una panoràmica general dels sectors socials involucrats en la recerca, el desenvolupament i l'ús de les tecnologies del català. Seguidament, a la secció 3, es plantegen una sèrie de qüestions polèmiques que afecten l'àmbit. Finalment, es presenten unes conclusions.

2. SECTORS SOCIALS IMPLICATS EN LES TECNOLOGIES LINGÜÍSTIQUES DEL CATALÀ

A més de la ciutadania, que n'és el destinari final, es poden distingir tres agents bàsics en l'àmbit de les tecnologies lingüístiques del català: la comunitat investigadora, l'empresa, i l'administració. Tots tres sectors en són alhora productors i usuaris, tot i que en mesures i de maneres diferents. En aquesta secció resumirem algunes dades sobre aquests tres sectors, tot i que ens centrarem en el primer.

Les dades sobre els grups de recerca que donarem a continuació estan basades en les presentacions que es van fer a la JPCC. Pensem que, tot i no ser exhaustives, són prou representatives de la recerca i el desenvolupament per al català de tecnologia lingüística. A l'annex presentem una taula que resumeix les característiques d'aquests grups.

Hi ha més d'una desena de centres de recerca a Catalunya i tres al País Valencià. A la resta de territoris de parla catalana (Illes Balears, Andorra, Catalunya Nord) no se n'hi coneixen, tot i que hi ha entitats que hi treballen com a mínim en el vessant dels recursos (Gabinet de Terminologia de la Conselleria d'Educació del Govern Balear) i empreses que fan servir algun tipus de tecnologia lingüística (IB3, El Periòdic Andorrà, etc.).

Encara en l'aspecte sociològic, cal observar que les tecnologies de la llengua són clarament interdisciplinàries, i s'hi dediquen esforços tant des de grups de recerca del món de les enginyeries (informàtica i telecomunicacions) com des de les humanitats (lingüística i traducció). Els grups presents a la JPCC mostren un equilibri considerable respecte aquest paràmetre, amb sis grups de les enginyeries i vuit de les humanitats.

Tots els grups treballen o bé en la recerca teòrica o en l'aplicada, i un d'ells específicament en la transferència de tecnologia, tot i que de diverses maneres tots els grups s'involucren en aquesta tasca. El finançament dels projectes és tant autonòmic com estatal i europeu i, a més dels concursos específics per a la recerca, la majoria dels grups tenen col·laboracions amb empreses i l'administració pública.

Pel que fa a aspectes tècnics, cal tenir en compte que una divisió bàsica en el tractament computacional de la llengua és la modalitat, oral o escrita. En general, hi ha més recerca en la modalitat escrita que en l'oral, i en el cas del català es corrobora aquesta tendència: deu grups tracten només text, tres s'ocupen de totes dues modalitats, i només un es dedica exclusivament a veu.

En la modalitat de veu les àrees bàsiques en què s'investiga són la síntesi o producció i el reconeixement de la parla. En la modalitat de text, la recerca bàsica es concentra en la creació d'eines de processament general (segmentadors, etiquetadors morfològics i sintàctics, etc.) i en la creació de recursos (gramàtiques i models, diccionaris, lèxics, corpus). Pràcticament tots els grups participen en major o menor mesura en les dues tasques, tot i que es tendeix a una divisió del treball entre enginyers (eines) i lingüistes (recursos).

Quant a les aplicacions, la traducció automàtica és el centre de la recerca: nou dels catorze grups hi dediquen esforços. Segueixen en l'interès la correcció i el resum automàtic (2 grups), i després taques diverses com ara l'extracció d'informació, la generació automàtica de textos o la detecció de paràfrasis.

Pel que fa a empreses productores de tecnologia lingüística que inclogui el català, en trobem de dedicades a la traducció automàtica (AutomaticTrans, Transducens, Translendum), a la correcció i revisió de textos (Barcelona Media, Maxigramar, Thera), a la recuperació i extracció d'informació (Barcelona Media, Inbenta, Thera), a la síntesi de parla (Barcelona Media, Locuendo, Telefónica I+D, Verbio), i al reconeixement de veu (Telefónica I+D, Verbio).

Així mateix, hi ha un bon nombre d'empreses i ens públics treballant en la creació i manteniment de recursos lingüístics, bàsicament lexicogràfics i terminològics (Thera, Assessorament Lingüístic de la Corporació Catalana de Mitjans Audiovisuals, Institut d'Estudis Catalans i TERMCAT), i alguns serveis lingüístics d'entitats administratives que treballen en l'explotació i la creació d'eines i recursos diversos com ara els multicercadors, les memòries de traducció i guies per a la redacció en català (a la Universitat Oberta de Catalunya, la Universitat de Barcelona, o la Comissió Jurídica Assessora de la Generalitat de Catalunya per citar-ne alguns).

Altres entitats que també col·laboren i contribueixen per aconseguir la normalitat del català en l'àmbit de la tecnologia, a més de les esmentades, són institucions com ara l'Acadèmia Valenciana de la Llengua o entitats del sector de la comunicació i la tecnologia com ara Softcatalà o VilaWeb. La implicació de tots aquests agents, com els hem anomenat al principi, es va poder constatar en les inscripcions de la JPCC, on al voltant d'un terç dels inscrits (de 166) provenia d'empreses, administració i d'altres entitats tant públiques com privades, tot i que la jornada estava dirigida principalment a la comunitat investigadora.

Gràcies a la feina de tot aquest col·lectiu, el català disposa de la majoria d'eines i aplicacions de tecnologia lingüística bàsiques per a qualsevol llengua, tot i que no totes són de lliure distribució ni tan sols per a propòsits de recerca (v. apartat següent). Podem destacar per exemple recursos com corpus anotats (p. ex., CTILC i CICA, desenvolupats per l'IEC), eines de processament genèriques (FreeLing, de la UPC, o LINLAP, de BM), o sistemes de traducció automàtica a diverses llengües (com Apertium o el traductor de Google).

En conjunt, doncs, podem dir que la situació de la tecnologia lingüística aplicada al català posa de manifest la varietat i l'abundància de productes tecnològics existents i la vitalitat de la comunitat desenvolupadora. Malgrat, tot, com veurem a la sessió dedicada a les qüestions controvertides al voltant d'aquest àmbit encara es percep una manca d'aprofitament dels recursos i una manca d'implicació per part de determinats àmbits de la societat civil.

3. QÜESTIONS CONTROVERTIDES A L'ÀMBIT DE LES TECNOLOGIES LINGÜÍSTIQUES DEL CATALÀ

3.1. Com es podria fer per ampliar la presència de la llengua catalana en els serveis i productes tecnològics? I per augmentar la transferència de la tecnologia que es desenvolupa als centres de recerca?

Existeix el consens generalitzat que la salut computacional del català és força bona; de fet, els recursos i eines disponibles són molts més que els que li correspondrien per demografia o poder polític, i la seva situació de gairebé normalitat en aquest sentit és excepcional per a una llengua sense estat.

Això no obstant, existeix una problemàtica específica que dificulta la difusió de les eines computacionals del català, tant per a objectius de recerca com per a aplicacions finals. Aquesta problemàtica es materialitza en els següents aspectes que tot seguit esmentem.

En primer lloc, la baixa demanda social de recursos tecnològics del català per part del món de l'empresa i els serveis (per exemple sectors com ara la banca o la sanitat) en què el català no és llengua vehicular real. Això fa que el desenvolupament privat de recursos en català sovint es faci per "voluntarisme". Tot i així, podria donar-se el cas que les pròpies tecnologies lingüístiques fessin el paper d'actor en la recuperació sociolingüística del català. Això succeiria si les eines en català fossin

prou útils i atractives. En aquest cas l'usuari les escolliria d'una manera espontània.

En segon lloc, la desconexió entre el desenvolupador i l'usuari final. En aquest sentit, des de certs àmbits s'apunta que les aplicacions distribuïdes amb llicències de codi obert (per exemple *Apertium*²) podrien facilitar aquesta connexió gràcies a les comunitats que s'estructuren al voltant d'aquest tipus de codi. Aquestes comunitats comparteixen i amplien les aplicacions segons els seus interessos i necessitats, i poden aglutinar tant grups de recerca, com usuaris i usuàries, o empreses orientades a serveis.

Un punt crucial és també la dificultat en l'accés a certes tecnologies i recursos per problemes de llicència. Aquí de nou les llicències de lliure distribució podrien tenir un paper important. En el cas que un recurs no es pogués fer de lliure distribució per problemes de propietat o *copyright*, com ara un corpus amb textos que són propietat d'autors o editorials, caldria facilitar com a mínim el seu ús per a objectius de recerca. En aquest cas, cal que les condicions d'ús estiguin ben especificades i que l'accés als recursos sigui àgil i sense obstacles burocràtics. Actualment, es dona una situació molt complicada, amb molta diversitat de llicències i condicions d'ús. Així, no és estrany el cas que per accedir a un recurs que és lliure per a recerca calgui esperar gairebé un any perquè cal signar un conveni entre les institucions implicades.

Especialment nocives són la duplicació i dispersió d'esforços en la construcció de corpus, eines, etc. El català és una llengua relativament petita, els recursos lingüístics són costosos i el finançament escàs. La impressió general és que no ens podem permetre com a comunitat malbaratar esforços d'aquesta manera. Cal doncs millorar la coordinació i fomentar la cooperació entre grups de recerca i també amb les empreses.

És també patent la relativament poca interacció entre lingüistes i informàtics a l'hora de crear i compartir aplicacions i recursos. Cal més interdisciplinarietat, i això es veu dificultat per les rígides estructures en l'organització acadèmica de la universitat, així com en la manca de centres de recerca híbrids.

Finalment, cal esmentar una certa indiferència per part dels investigadors en tecnologies del llenguatge per la qüestió de la llengua. Això fa que llengües com l'anglès, que són les que compten amb més recursos, actuïn com a atractors dels esforços investigadors, desequilibrant cada vegada més la proporció de recursos entre les llengües. Per combatre aquest fenomen, cal afavorir la presència de recursos textuais per al català, com ara corpus anotats, en competicions internacionals per a tasques computacionals.

3.2. Quines estratègies de finançament són necessàries? Quina relació hi ha d'haver entre l'origen del finançament i les condicions d'accessibilitat i distribució dels recursos?

Hi ha l'opinió extesa que calen polítiques d'incentivació de la innovació a les empreses. Les polítiques actuals, però, estan orientades a projectes massa grossos només assumibles per grans empreses, com Microsoft. Les accions de l'administració podrien ser més efectives si es fessin polítiques orientades a PIMES.

Així mateix, per tal d'afavorir el codi obert, cal que en les convocatòries de projectes competitius, les entitats que solliciten els ajuts hagin d'especificar les condicions d'ús dels recursos que es generaran, per tal que els avaluadors ho puguin tenir en compte a l'hora de concedir el projecte i de controlar-ne el progrés. Aquesta demanda es podria fer a nivell estatal a fi d'incorporar-la a la nova Llei de la Ciència.

Una qüestió afegida en relació a la lliure distribució dels recursos generats pels projectes de recerca és que cal contemplar els costos de l'empaquetament final de l'aplicació. Aquests costos, així com la preparació de llicències i el mecanisme de distribució, s'haurien de preveure en la definició inicial dels projectes.

2 <http://xixona.dlsi.ua.es/apertium>

3.3. Com s'hauria d'articular la comunitat que formen les persones vinculades d'una manera o altra a l'àmbit del Processament del Llenguatge Natural i de la Parla de la llengua catalana?

Seguint el model de la Linguateca³ de la llengua portuguesa, es podria crear un portal web de les tecnologies lingüístiques del català que centralitzés tota la informació referent als recursos, les eines, els grups i empreses que hi treballen, etc. Respecte al suport necessari per crear el portal i sobretot mantenir-lo, hi ha diferents alternatives. Es pot aprofitar el marc d'algun projecte ja existent, ja sigui en l'àmbit de l'administració local o estatal (per exemple amb el programa AVANZA), per tal que els poders públics financin aquest manteniment. També hi ha la possibilitat alternativa de crear-lo col·lectivament en forma de wiki cooperativa. Existeixen també iniciatives properes com el projecte europeu CLARIN, que ja inclou el català, i que pretén facilitar l'explotació de recursos lingüístics per part dels investigadors en ciències socials.

Per tal de possibilitar iniciatives com la del portal web, fora interessant comptar amb una associació que donés estructura a la comunitat vinculada amb l'àmbit de les tecnologies del català, a l'estil de la francesa (ATALA) o l'espanyola (SEPLN). Una associació permetria donar veu al col·lectiu de recerca i alhora fer de nexa tant amb les empreses com amb l'administració i els usuaris i usuàries finals.

4. CONCLUSIONS

Atesa la quantitat i qualitat de grups de recerca i d'altres entitats públiques i privades, així com els productes, eines i recursos existents és evident la vitalitat tecnològica del català. Cal però incidir en alguns aspectes que dificulten la transferència tecnològica entre grups, i entre la recerca i el desenvolupament de productes i serveis.

Pel que fa a la col·laboració i l'aprofitament dels recursos cal una acció comuna i activa per part de la comunitat, però també suport institucional. En particular, per establir una política més oberta de distribució i una major facilitat d'ús de la tecnologia lingüística cal un suport decidit dels poders públics, tan autonòmics com estatals i europeus, ja que són ells els que estableixen les condicions de finançament dels projectes de recerca i de transferència de tecnologia.

És necessari també articular iniciatives per intensificar la comunicació i la col·laboració entre els diferents agents de la comunitat tecnològica catalana, tant entre els mateixos grups de recerca com amb l'empresa, ja sigui desenvolupadora o usuària. La Jornada del Processament Computacional del Català ha estat una iniciativa en aquesta línia, però cal anar més enllà amb la creació d'un portal web que centralitzi les eines i els recursos i idealment amb la constitució d'una associació que aglutini tota la comunitat implicada.

ANNEX

Grup de Veü i Llenguatge (Barcelona Media Centre d'Innovació)
Objectiu: el desenvolupament de nous productes i serveis, i la potenciació de la recerca i el desenvolupament (R+D).
Línies de treball: extracció d'informació, correcció i revisió de textos, traducció automàtica, síntesi de veü, llengua de signes, i tecnologies per a la docència de la traducció i la lingüística.
URL: http://www.barcelonamedia.org/linies/7/ca
Grup de Processament del Llenguatge Natural, Centre de Tecnologies i Aplicacions del Llenguatge i la Parla (Universitat Politècnica de Catalunya)
Objectiu: la recerca en el camp del processament computacional del llenguatge natural, així

3 <http://www.linguateca.pt/>

com la creació de recursos lingüístics.

Línies de treball: recerca en processadors lingüístics, adquisició, integració i explotació de coneixement lexico-semàntic, anàlisi semàntica des de diferents punts de vista (WSD, SRL, NLU), aprenentatge automàtic aplicat al PLN, resum automàtic i traducció automàtica.

URL: <http://www.lsi.upc.edu/~nlp>

Grup de Tractament Automàtic del Llenguatge Natural,
Dept. Tecnologies de la Informació i les Comunicacions
(Universitat Pompeu Fabra i Barcelona Media Centre d'Innovació)

Objectiu: fer recerca i aplicacions en alguns camps específics del processament automàtic del llenguatge natural, en el marc teòric de la Teoria Sentit-Text (TST).

Línies de treball: generació multilingüe de llenguatge natural i altres continguts, resum automàtic, paràfrasi, traducció, lexicologia computacional, i aprenentatge automàtic orientat a l'adquisició de recursos lingüístics.

URL: <http://www.recerca.upf.edu/taln>

Servei Lingüístic (Universitat Oberta de Catalunya)

Objectiu: fer un ús intensiu de la tecnologia disponible per tal d'optimitzar els processos productius (rendibilitzar costos i aportar qualitat i homogeneïtat).

Línies de treball: l'explotació i la millora dels sistemes de Traducció Automàtica, i la creació d'eines i recursos de suport als processos de correcció i traducció.

URL: <http://www.uoc.edu/serveilinguistic>

FlexSem,

Departament de Filologia Francesa i Romànica (Universitat Autònoma de Barcelona)

Objectiu: la recerca aplicada al processament automàtic del llenguatge natural des d'una perspectiva de formalització lingüística i la creació de recursos lingüístics.

Línies de treball: prosòdia, patologies de la parla i intercomprensió entre llengües romàniques; les vinculacions i vehiculacions semàntiques del lèxic amb aplicabilitat en la creació de diccionaris electrònics i mòduls de processament lingüístic.

URL: <https://masters.uab.es/flexsem>

Centre de Llenguatge i Computació (Universitat de Barcelona)

Objectiu: desenvolupar recursos de tecnologia lingüística (corpus anotats, lexicons, analitzadors) que són a la base de les aplicacions basades en processament del llenguatge, amb cura dels fonaments lingüístics i metodològics.

Línies de treball: processament massiu de textos, anàlisi morfològica i sintàctica, corpus anotats, analitzadors basats en ML, resolució de la coreferència, identificació de paràfrasis.

URL: <http://clic.ub.edu>

Institut Universitari de Lingüística Aplicada (Universitat Pompeu Fabra)

Objectiu: la creació de recursos lingüístics i la recerca basada en la lingüística de corpus i les tecnologies lingüístiques.

Línies de treball: demolingüística, discurs especialitzat, documentació digital, enginyeria lingüística, extracció d'informació, fonologia aplicada, lexicografia, lexicologia, lingüística forense, neologia, sociolingüística, terminologia i variació lingüística.

URL: <http://www.iula.upf.edu>

Grup de Recerca en Tecnologies Mèdia Enginyeria i Arquitectura la Salle
(Universitat Ramon Llull)

Objectiu: treballar la multidisciplinarietat centrada en la innovació en noves tecnologies multimèdia i multimodals.

Línies de treball: síntesi de la parla expressiva, creació d'avatars (cares parlants i LSC), reconeixement de la parla, reconeixement audiovisual d'emocions i finalment, disseny i enregistrament de corpus de veu.

URL: <http://www.salle.url.edu/~iriondo>

Grup de Recerca Interuniversitari en Aplicacions Lingüístiques,
Grup interuniversitari (Universitat Autònoma de Barcelona, Universitat de Barcelona i Universitat Oberta de Catalunya)

Objectiu: desenvolupar, avaluar i aplicar recursos lingüístics de gran cobertura, creant i fent ús de gramàtiques, lèxics, corpus i memòries de traducció.

Línies de treball: gramàtiques (dependències, HPSG), recursos lèxics, corpus (anotació i explotació), adquisició d'informació (esquemes de subcategorització i restriccions de selecció) i memòries de traducció.

URL: <http://grial.uab.es>

Pattern Recognition and Human Language Technology,
Institut Tecnològic d'Informàtica (Universitat Politècnica de València)

Objectiu: el reconeixement de patrons, l'aprenentatge automàtic i la interacció multimodal.

Línies de treball: traducció de textos i veu (automàtica o assistida), reconeixement de la parla, reconeixement de texts manuscrits (automàtic o assistit), biometria i recuperació d'imatges interactiva.

URL: <http://prhlt.iti.es>

Grup de Tractament de la Parla,
Centre de Tecnologies i Aplicacions del Llenguatge i la Parla
(Universitat Politècnica de Catalunya)

Objectiu: la recerca en el camp del processament computacional de la llengua així com en la creació de recursos lingüístics, amb un interès especial cap a la traducció automàtica.

Línies de treball: reconeixement automàtic de la parla, conversió de text a veu, traducció automàtica oral i textual, creació de recursos lingüístics (orals i textuals) i el processat de la parla i àudio en interfícies multimodals.

URL: <http://gps-tsc.upc.es/veu>

Grup Transmedia Catalonia,
Dept. de Traducció i d'Interpretació (Universitat Autònoma de Barcelona)

Objectiu: l'experimentació mitjançant tecnologies com l'*eye-tracking* o el reconeixement de parla aplicades a la traducció audiovisual i a l'accessibilitat.

Línies de treball: traducció audiovisual (doblatge, subtitulació, veus superposades), accessibilitat als mitjans (audiodescripció, subtitulació per a sords i audiosubtitulació) i accessibilitat en la docència.

URL: <http://www.fti.uab.cat/transmediacatalonia>

Projecte institucional d'investigació IVITRA (Universitat d'Alacant)

Objectiu: la recerca d'eines relacionades amb la traducció automàtica i d'aplicacions per al tractament de corpus multilingües i paral·lels.

Línies de treball: elaboració de corpus del català antic, textos paral·lels (traduïts) i totes les eines relacionades amb el seu tractament.

URL: <http://www.ivitra.ua.es>

Transducens,
Dept. de Llenguatges i Sistemes Informàtics (Universitat d'Alacant)

Objectiu: la recerca i la creació d'aplicacions que emprin tecnologies de processament del

llenguatge natural.

Línies de treball: l'educació assistida per ordinador, la indexació i llenguatges de marcat en biblioteques digitals, i la inferència gramatical a partir de mostres estocàstiques i les seves aplicacions; traducció automàtica, anotació fonètica de textos per a la lectura en veu alta, i eines per al processament de textos paral·lels i generació de memòries de traducció.

URL: <http://transducens.dlsi.ua.es>