Llengua i Ús
Revista Tècnica de Política Lingüística
58
1r SEMESTRE 2016
ISSN: 2013-052X
http://gencat.cat/llengua/liu

Models
1.2

Les indústries de la llengua i la tecnologia per al català

Es indústries dera lengua e era tecnologia entath catalan

Las industrias de la lengua y la tecnología para el catalán

The Language Industries and Technology for Catalan

Núria Bel i Montserrat Marimon
Universitat Pompeu Fabra

Citació recomanada:
Bel, Núria; Marimon, Montserrat. «Les indústries de la llengua i la tecnologia per al català». *Llengua i Ús: Revista Tècnica de Política Lingüística* [en línia] [Barcelona: Generalitat de Catalunya. Departament de Cultura. Direcció General de Política Lingüística], núm. 58 (1r semestre 2016), p.?.<http://www.raco.cat/index.php/LlenguaUs/article/view/311744/401822>

El català *Llengua per a tothom*

Generalitat de Catalunya
Departament de Cultura

Les indústries de la llengua i la tecnologia per al català

Tecnologies del llenguatge,processament del llenguatge natural, indústries de la llengua, traducció automàtica, recursos lingüístics, infraestructures

L'estiu del 2015 es va presentar l'*Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España*. En aquest informe un ampli grup d'experts va fer una anàlisi de les fortaleses, debilitats, oportunitats i amenaces de la situació de les indústries de la llengua en el seu vessant tecnològic. En aquest article reprenem aquesta anàlisi fent èmfasi en la situació a Catalunya per identificar els reptes específics que s'han de superar perquè la indústria produeixi aplicacions per al català. Com la majoria de les llengües europees, el català s'enfronta al repte de decidir si ha de compensar la mida del seu mercat potencial amb estratègies de suport públic, una de les quals pot ser la creació d'infraestructures lingüístiques.

Es indústries dera lengua e era tecnologia entath catalan

tecnologies deth lenguatge, tractament deth lenguatge naturau, indústries dera lengua, traduccion automatica, recorsi lingüistics, infrastructures

En estiu de 2015 se presentèc er *Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España*. En aqueth arrepòrt un vast grop d'expèrts hec ua analisi des fortaleses, des febleses, des escadences e des menaces dera situacion des indústries dera lengua en sòn encastre tecnologic. En aguest article reprenem aquera analisi en tot hèr enfasi ena situacion en Catalonha entà identificar es enjòcs especifics que cau superar entà qu'era indústria produsisque aplicacions entath catalan. Coma era majoritat des lengües europèes, eth catalan afronte er enjòc de decidir s'a de compensar era mesura deth sòn mercat potenciau damb estrategies de sosten public, qu'ua pòt èster era creacion d'infrastructures lingüistiques.

Las industrias de la lengua y la tecnología para el catalán

tecnologías del lenguaje, procesamiento del lenguaje natural, industrias de la lengua, traducción automática, recursos lingüísticos, infraestructuras

El verano de 2015 se presentó el *Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España*. En este informe un amplio grupo de expertos hizo un análisis de las fortalezas, debilidades, oportunidades y amenazas de la situación de las industrias de la lengua en su vertiente tecnológica. En este artículo retomamos dicho análisis haciendo énfasis en la situación en Cataluña para identificar los retos específicos que se deben superar para que la industria produzca aplicaciones para el catalán. Como la mayoría de las lenguas europeas, el catalán se enfrenta al reto de decidir si debe compensar el tamaño de su mercado potencial con estrategias de apoyo público, una de las cuales puede ser la creación de infraestructuras lingüísticas.

The Language Industries and Technology for Catalan

language technology, natural language processing, language industries, machine translation, language resources, infrastructure

In the summer of 2015 the *Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España* (Report on the State of Language Technologies in Spain within the Digital Agenda for Spain) was presented. In this report a large group of experts examined the strengths, weaknesses, opportunities and threats of the situation of the language industries in their technological aspect. In this paper we continue this analysis with emphasis on the situation in Catalonia to identify the specific challenges that need to be overcome so that the industry can produce applications for Catalan. Like most European languages, Catalan faces the challenge of deciding whether to make up for the size of its potential market with public support strategies, one of which may be setting up language infrastructures.

Introducció

Aplicacions com la traducció automàtica (TA), els assistents virtuals, els correctors gramaticals automàtics i la classificació automàtica de documents són ara ja eines populars i les grans consultores pronostiquen que en els pròxims cinc anys les tecnologies del llenguatge, el Processament del Llenguatge Natural (PLN), seran la base de l'aparició de moltes noves aplicacions que llegiran, entendran i extrauran informació en àmbits molt diversos. El PLN està considerat una tecnologia clau del processament de dades massives ja que permet extreure dades útils, per exemple, de les més de cent mil piulades per minut que es fan a tot el món (Paniagua, 2013) per: analitzar la qualitat dels fàrmacs a partir de l'opinió dels consumidors (Las redes sociales, 2015), predir èxits editorials (Generalitat de Catalunya, 2016) o preveure el comportament de la borsa (Sprenger *et al*., 2014); o dels comentaris dels 60 milions d'informes de salut disponibles digitalment a Catalunya (Generalitat de Catalunya, 2015) per trobar interaccions entre fàrmacs.

L'estiu de 2015 es va presentar l'*Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España* (Bel i Rigau, 2015). En aquest informe un grup d'experts va revisar la situació a Espanya de les tecnologies del llenguatge. L'*Informe* va fer una anàlisi FODA (Fortaleses, Debilitats, Oportunitats i Amenaces) després d'estudiar la disponibilitat de recursos lingüístics en espanyol i llengües cooficials, la situació de la recerca a Espanya, el mercat i la indústria de la llengua per a l'espanyol i llengües cooficials i finalment l'interès d'aquestes aplicacions per a l'Administració pública.

En aquest article repassem aquesta anàlisi fent èmfasi en la situació a Catalunya per identificar els reptes que el català ha de superar en els propers anys per tal que disposi d'aplicacions que processin també aquesta llengua.

L'Estat de les tecnologies de la llengua a Catalunya, una anàlisi FODA

Una anàlisi FODA es basa en la identificació de fortaleses, debilitats, oportunitats i amenaces que presenta una situació determinada amb l'objectiu d'identificar reptes i estratègies que permetin millorar-la i optimitzar-la.

Quant a fortaleses, l'informe mostra que Catalunya disposa de capacitat tecnològica i experiència provada en les tecnologies del llenguatge. El servei gratuït de TA que la Generalitat de Catalunya va posar en marxa del 2006 al 2010, el sistema especialitzat en l'àmbit jurídic que dóna serveis als professionals de l'Administració de justícia i l'ús de TA en la publicació bilingüe de premsa (*El Segre* o *El Periódico* des dels anys noranta) mostren clarament que a Catalunya es va veure aviat el paper clau d'aquesta tecnologia. El clúster català d'indústries de la llengua, ClusterLingua, creat el 2011, mostra l'existència d'un sector econòmic i engloba empreses que desenvolupen eines per al tractament automàtic de la llengua. Les empreses del sector, unes vint d'identificades, són tecnològicament capdavanteres, algunes amb presència internacional, en particular en el reconeixement de la parla i la TA.

També són considerats capdavanters i de prestigi internacional els deu grups de recerca, repartits a totes les universitats catalanes, i amb projectes de col·laboració amb empreses, projectes de recerca i innovació subvencionats pel Programa marc de la Comissió de la Unió Europea o el Pla nacional d'investigació i desenvolupament propi de tecnologia. En particular destaca el processador FreeLing (Padró i Stanilovsky, 2012), programa de codi obert amb més de 250.000 descàrregues des del 2009.

El PLN està vivint un *momentum* al món i l'informe ho veu com una oportunitat. Les grans consultores (Gartner, 2014) el relacionen amb el creixement d'aplicacions d'anàlisi de les dades, en forma de text, de les xarxes socials i de l'Administració pública que la Llei 37/2007, de

16 de novembre, sobre reutilització de la informació del sector públic, promociona amb portals de dades obertes. Per altra banda, grans multinacionals han popularitzat algunes aplicacions, per exemple, la TA (Google i Microsoft tenen sengles productes gratuïts), els assistents virtuals (com Siri d'Apple) i la resposta automàtica a preguntes (Watson d'IBM).

A Europa, diferents iniciatives promouen el multilingüisme de les aplicacions del PLN, ara majoritàriament disponibles solament per a l'anglès. S'han creat associacions com la Multilingual Europe Technology Alliance i The Language Technology Industry Association i s'estan posant en marxa infraestructures com la Connecting Europe Facility, Automated Translation (CEF-AT) i iniciatives com l'European Language Resource Coordination (ELRC), que proposen la utilització oberta dels serveis de TA de la Comissió Europea.

A Espanya, el Plan de impulso de las tecnologías del lenguaje, de la Secretaria de Estado de Telecomunicaciones y para la Sociedad de la Información (SETSI), creat en el marc de l'Agenda Digital per a Espanya, va ser presentat l'octubre del 2015, amb un pressupost de 89 milions d'euros, amb l'objectiu d'augmentar el nombre, la qualitat i la disponibilitat d'aplicacions per al castellà, el català, el basc i el gallec i donar impuls a la indústria del llenguatge i la seva internacionalització, i té previst usar la contractació pública per proveir l'Administració (també les de les comunitats autònomes) d'aplicacions innovadores.

L'informe considera una debilitat crucial la manca de visibilitat que pateix el sector en el context mundial, en particular pel desconeixement que, tot i que la tecnologia actual pot processar textos en qualsevol llengua, per fer-ho necessita informació específica de la llengua dels textos que ha de processar. La carència de dades lingüístiques, o de recursos lingüístics, d'una llengua en particular significa que no hi pot haver aplicacions per a aquesta llengua. Els informes que META-NET va publicar el 2012 (Moreno *et al*., 2012) mostren que la disponibilitat de recursos lingüístics per a les llengües europees és desigual i situen el català al mateix nivell del neerlandès, l'alemany, l'hongarès, l'italià, el polonès i el romanès, però clarament per sota de l'anglès, l'espanyol i el francès.

Aquests recursos lingüístics, la gasolina dels motors de les aplicacions, són textos, documents, diccionaris electrònics amb informació lingüística i els processadors lingüístics que se'n deriven i, des d'aquest punt de vista, són una infraestructura necessària per al desenvolupament d'aplicacions. La manca de recursos disponibles per a ús comercial és especialment crítica per a les petites empreses, característiques a Catalunya, perquè, tot i que siguin capdavanteres en tecnologia, el desenvolupament de recursos lingüístics, que es converteixen en un factor competitiu, acapara la seva capacitat d'inversió i les manté limitades a un mercat reduït: afegir una nova llengua o canviar d'àmbit (per exemple, passar una aplicació sobre textos legals a textos clínics) suposa invertir en la creació de nous recursos. A més, el cost de fer recursos fa impossible l'ús d'aquestes tecnologies per part d'empreses emergents (o *start-ups*), que són, segons les grans consultores, l'origen de les aplicacions innovadores i d'èxit que han d'aparèixer en els propers anys.

En l'apartat d'amenaces, l'informe assenyala que el desenvolupament d'aplicacions per a una llengua està condicionat per les dimensions del segment de mercat que representa el nombre de parlants de la llengua en qüestió. Un mercat com l'europeu, tan fragmentat lingüísticament, no és atractiu comercialment, i no es produeixen aplicacions per a moltes de les seves llengües. La situació sociopolítica del català i les dimensions del seu mercat apunten a una progressiva extinció digital, com anunciava l'informe de META-NET: el risc és que aquestes aplicacions no arribin a tractar textos en català.

De l'anàlisi de les debilitats i de les oportunitats podem extreure conclusions sobre els reptes que el català ha de superar per no «morir digitalment». Ja hem dit que la de-

bilitat principal és la manca d'una infraestructura lingüística. L'oportunitat és que aquestes aplicacions són cada cop més populars i es constata la seva mancança per a llengües europees, els parlants de les quals ja comencen a reclamar la creació d'aquesta infraestructura perquè les pimes i les empreses emergents que vulguin desenvolupar aplicacions puguin reduir la inversió important que suposa la creació de recursos lingüístics propis. El Plan de impulso de la SETSI inclou, entre les seves mesures, inversió pública per a la creació d'infraestructures lingüístiques, tanmateix és competència de la Generalitat de Catalunya dissenyar i constituir la infraestructura per al català.

La reutilització de dades públiques és una forma efectiva de contribuir a la creació d'aquesta infraestructura com mostren aquests exemples: La col·lecció de textos i traduccions (les memòries de traducció), que són la base dels sistemes de TA estadística de la Direcció General de Traducció de la Comissió Europea, es poden trobar al Portal Europeu de Dades Obertes (i és el paquet més descarregat d'entre els més de 8.000 paquets de dades de tota mena que s'hi ofereixen), així com les memòries de traducció de l'Escola d'Administració Pública del Govern Basc (IVAP) i les dades terminològiques del TERMCAT al seus portals respectius.

Així doncs, identificar i publicar documents, glossaris i thesaurus com a dades obertes és una solució, efectiva i ja abastable, per contribuir a la infraestructura lingüística del català. Però hi ha algunes qüestions tècniques importants que cal tenir en compte. Les tres més importants són:

- Primer, els documents i les traduccions, els glossaris i els thesaurus són recursos molt valuosos que poden ser reutilitzats, però han de ser subministrats en uns formats electrònics processables (especialment que no sigui PDF).

- Segon, s'hi ha de tenir accés per descàrrega de fitxers. Molts organismes ofereixen dades mitjançant l'accés a través d'una aplicació web de cerca, però no permeten la descàrrega de fitxers.

- Tercer, les dades s'han d'oferir amb una llicència que en permeti la còpia, la transformació i la creació de derivats perquè s'ha de preveure que la indústria les pugui fer servir.

Amb la contribució dels textos de forma massiva i constant per part d'organismes públics es resoldria una part de la infraestructura. Hi ha recursos lingüístics necessaris per a aplicacions concretes, la creació dels quals requereix treballs específics i una inversió que, tot i que en els darrers anys ha estat pobra i discontínua, ha fet que el català, com hem vist, estigui al mateix nivell que altres llengües europees. Però els recursos disponibles són encara limitats i disten molt, en nombre i tipus de tecnologia coberta, dels disponibles per a l'espanyol i, evidentment, per a l'anglès. A continuació fem una revisió dels disponibles actualment i n'avaluem les mancances.

Infraestructura lingüística per al català

Els principals productors de recursos lingüístics disponibles per al català han estat els grups d'investigació de diferents universitats catalanes, tot i que també l'Institut d'Estudis Catalans (IEC) i el TERMCAT han posat a disposició d'aplicacions les seves dades lingüístiques en diferents ocasions. La majoria s'han desenvolupat amb finançament públic. La Generalitat de Catalunya va finançar el Centre de Referència d'Enginyeria Lingüística (1996-2000), coordinat per l'IEC i amb la participació de diferents universitats catalanes, amb la missió de promoure la creació d'eines i recursos per al processament automàtic de textos en català. El Pla nacional d'investigació i tecnologia ha estat també una font de finançament, així com el Programa marc de R+D europeu. Malgrat aquest esforç, és important fer notar que molts no poden formar part d'una infraestructura lingüística per a

ús industrial, ja que es distribueixen amb llicències per a usos no comercials.

Corpus textuals

Són col·leccions de textos etiquetats amb diferent informació gramatical que pot anar des d'informació morfosintàctica a informació de constituents i funcions sintàctiques, estructura argumental i papers temàtics, classes semàntiques i sentits, i entitats amb nom i relacions de coreferència. Aquests recursos bàsics són la base per crear els diferents processadors lingüístics.

El **Corpus Textual Informatitzat de la Llengua Catalana** (Rafel, 1994), creat per l'IEC, només està disponible per a la recerca. Consta de 52 milions de paraules de textos escrits entre el 1832 i el 1988. En canvi, el corpus català PAROLE, de 21 milions de paraules, també de l'IEC i creat amb finançament europeu, tot i no tenir cap llicència específica, es cedeix mitjançant conveni també per a ús industrial.

El **Corpus Tècnic de l'IULA**, desenvolupat per l'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra (UPF) (Cabré *et al*., 2006) i amb llicència gratuïta per a recerca, conté textos tècnics en català de 23 milions de paraules lematitzades i anotades amb informació morfosintàctica automàticament i està sent actualment anotat amb informació sintàctica de dependències.

Els **Ancora-CA** i **AnCora_DEP-CA**, del Centre de Llenguatge i Computació (CLIC) de la Universitat de Barcelona (UB) i el Centre de Tecnologies i Aplicacions del Llenguatge i de la Parla (TALP) de la Universitat Politècnica de Catalunya (UPC), tenen llicència de codi obert GPL que pot no ser adequada per a usos industrials ja que obliga que el programa del qual formen part sigui també de codi obert i no totes les empreses hi estan interessades. Consta de 500.000 paraules de textos de premsa anotats amb informació morfosintàctica, sintàctica i semàntica (Taulé *et al*., 2008).

El **Wikicorpus**, un corpus trilingüe (català, espanyol, anglès) constituït amb una part de la Wikipedia està anotat de forma automàtica amb informació morfològica (Reese *et al*., 2010). La part catalana, de 50 milions de paraules, va ser creada pel TALP i té una llicència GNU FDL, la mateixa que fa servir la Wikipedia per donar els drets d'ús, modificació i distribució dels seus textos, però amb l'obligació que el resultat tingui també aquesta llicència.

Corpus paral·lels

Aquests corpus es componen de textos originals i les seves traduccions alineats per segments, o unitats de traducció. Són útils com a memòries de traducció, per a tasques d'extracció automàtica de terminologia i són la base dels sistemes de TA estadística.

El **Corpus paral·lel català-castellà** del *Diari Oficial de la Generalitat de Catalunya* inclou 715.000 documents (del 1977 fins al 2015) i es pot descarregar des del web del grup de recerca Language Processing Group de la Universitat Oberta de Catalunya amb llicència oberta.

El **Corpus paral·lel espanyol-català** de més de 100 milions de paraules de 10 anys d'articles d'*El Periódico de Catalunya* el distribueix l'European Language Resources Association amb una llicència que en permet l'ús comercial.

Recursos lèxics i terminològics

Els diccionaris i les terminologies són llistes de paraules amb informació lingüística (morfosintàctica, sintàctica i semàntica) necessaris per a diferents tipus de processadors lingüístics que extreuen informació sobre participants en accions i relacions entre ells.

El **Lèxic morfològic del català** del Corrector, utilitzat pel corrector gramatical desenvolupat pel Grup de Lingüística Computacional (GLiCom) de la UPF (Badia *et al*., 2009) inclou 640.000 entrades (71.000 lemes) i es distribueix amb llicència GNU GPL.

El **Lèxic morfològic del català** d'Apertium (Forcada *et al*., 2011), amb 11.800 lemes amb informació morfològica i els diccionaris bilingües *Apertium*, de català amb anglès, francès, italià, portuguès, espanyol, occità i esperanto, que varien per parell de llengües: entre les 9.000 i les 33.000 entrades. Són descarregables al web del grup Transducens de la Universitat d'Alacant i es distribueixen sota llicència GNU.

La **Base de dades SENSEM**, creada pel Grup de Recerca Interuniversitari en Aplicacions Lingüístiques (GRIAL), té llicència GPL, també. Desenvolupat a partir del corpus SENSEM, que conté 100 frases per a cada un dels 250 verbs més freqüents de l'espanyol actual traduïts al català (Vázquez i Fernández, 2011), inclou estructura argumental, patrons de subcategorització (amb informació de freqüència), papers semàntics i informació semàntica de l'oració per a cada verb.

El **Lèxic PAROLE-SIMPLE** del català de 20.000 entrades amb informació morfològica, sintàctica i semàntica segons l'estàndard EAGLES va ser desenvolupat per l'IEC i es distribueix amb llicència Creative Commons (CC-BY-NC-SA), no apte per a ús comercial.

Els lèxics amb informació semàntica més usats al món són els WordNets. Per al català la darrera versió és WordNet 3.0, amb 46.442 *synsets* que forma part del Repositori Central Multilingüe, que integra també els de l'espanyol, el basc, el gallec i l'anglès (González-Agirre, *et al*., 2012), descarregable i amb llicència CC-BY.

El **Banc de neologismes**, de l'Observatori de Neologia de la UPF, és un recull dels neologismes procedents de la premsa catalana i castellana des de l'any 1992. Es distribueix sota llicència Creative Commons (CC-BY-NC-SA), no comercial.

El **Lèxic fonètic del català LC-STAR**, desenvolupat pel grup TALP (Bisami *et al*., 2003), consta de 100.000 paraules, distribuïdes en tres categories: 53.225 paraules extretes d'un corpus de més de 20 milions de paraules, 45.306 noms propis i 7.498 paraules traduïdes de termes en anglès. També distribuït per ELRA, té llicència compatible amb l'ús industrial.

Pel que fa als recursos terminològics, a més dels oferts pel TERMCAT, els glossaris i la terminologia dels productes informàtics localitzats per Softcatalà es poden descarregar al seu web amb llicència CC-BY-SA. Finalment, Microsoft ofereix de forma gratuïta la terminologia dels seus productes en català i Apple n'ofereix la seva als desenvolupadors d'Apple registrats.

Corpus Orals

Els corpus orals són bases de dades d'arxius de veu i les seves transcripcions que en la tecnologia de la parla s'usen per crear motors de reconeixement de veu i síntesi de la parla.

El grup TALP i l'empresa Verbio han compilat diversos corpus orals en el marc de diferents projectes i són distribuïts per ELRA i amb llicències per a usos comercials. Llistem a continuació les seves característiques.

- **SpeechDat CAR Catalan**, gravacions mitjançant quatre micròfons instal·lats en cotxes de 600 sessions de 300 informants de 5 regions dialectals diferents i equilibrats per sexe i edat. **SpeechDat Catalan FDB 2000 speakers** i **SpeechDat Catalan MDB 2000 speakers**, enregistraments de 2.000 informants, d'ambdós sexes, registrades des de telèfons fixos i mòbils. Contenen transcripcions ortogràfiques i informació de l'edat i regió dialectal dels informants. **SpeechDat Catalan FDB 1005 speakers**: enregistraments de 1.005 informants d'ambdós sexes i diferents edats enregistrades des de telèfons fixos.

- **FESTCAT Catalan TTS baseline male/female speech database**, enregistraments i anotacions de material de text llegit, d'aproximadament 20 hores, i **FESTCAT-SEL**: enregistraments i anotacions de 8 informants de material de text llegit d'aproximadament una hora.

- **SpeechCon Catalan**, enregistraments de 550 informants, d'ambdós sexes, de locucions llegides i parla espontània enregistrades amb 4 micròfons mitjançant una plataforma mòbil. Altres de característiques semblants són: **TALP Tourism Dialogues**, 22 hores de gravacions de diàlegs en el domini turístic. **TM2**, que conté els enregistraments de dues reunions tècniques en espanyol i català, **3/24 BN (Catalan BN)**, de 80 hores de notícies; **AGORA**, gravacions de 34 programes de TV3 amb dades segmentades i anotades.

Finalment, el corpus **Glissando**, anotat per a estudis prosòdics, inclou els enregistraments de 40 hores de 28 locutors, professionals i no professionals (Garrido, 2012) i ha estat desenvolupat pel GLiCom-UPF, el Grup d'Estudis en Prosòdia de la UAB i el grup Sistemes d'Interacció Multimodal de la Universitat de Valladolid.

Conclusions

Si parlar i escriure documents i llegir-los per extreure'n informació forma part de gairebé totes les professions, llavors el PLN és un component que permet optimitzar el treball i reduir costos en gairebé tots els sectors d'activitat. Així doncs, en els propers anys hem de veure moltes aplicacions innovadores i útils de les tecnologies del llenguatge. Però, com la majoria de les llengües europees, el català s'enfronta al repte de compensar la mida del seu mercat potencial amb estratègies de suport que redueixin la inversió que haurien de fer les empreses en la creació dels recursos lingüístics necessaris. És urgent avaluar la conveniència de posar a l'abast de la indústria del llenguatge una infraestructura lingüística creada amb suport públic.

Bibliografia

Badia, Toni [*et al*.] (2009). «Un corrector tipogràfic, ortogràfic i gramatical de català». 1a Jornada del Processament Computacional del Català. Barcelona.

Bel, Núria; Rigau, German (ed.) (2015). *Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España* [en línia]. <http://www.agendadigital.gob.es/planes-actuaciones/tecnologias-lenguaje/Bibliotecaimpulsotecnologiaslenguaje/Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf> [Consulta: 28 juny 2016].

Benítez, Laura [*et al*.] (1998). «Methods and tools for building the Catalan WordNet». A: *Proceeding of LREC-1998* [en línia]. Granada: Universitat de Granada. <http://www.cs.upc.edu/~escudero/wsd/98-lrec.pdf> [Consulta: 28 juny 2016].

Bisami, Maximilian [*et al*.] (2003). «Lexicon and Corpora for Speech to Speech Translatoin (LC-STAR)». *Procesamiento del lenguaje natural*, núm. 31, p. 317-318. També disponible en línia a: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/3189/1680> [Consulta: 28 juny 2016].

Boleda, Gemma [*et al*.] (2006). «CUCWeb: a Catalan corpus built from the Web». A: *Proceedings of Second Workshop on the Web as a Corpus at EACL'06* [en línia]. Trento, Itàlia. <http://www.aclweb.org/anthology/W06-1704> [Consulta: 28 juny 2016].

Cabré, M. Teresa; Bach, Carme; Vivaldi, Jorge (2006) [en línia]. *10 anys del Corpus de l'IULA*. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra. (Papers de l'IULA; Informes, 44) <http://repositori.upf.edu/bitstream/handle/10230/1298/06inf044.pdf?sequence=1> [Consulta: 28 juny 2016].

Forcada, Mikel L. [*et al*.] (2011). «Apertium: a free/open-source platform for rule-based machine translation». *Machine Translation* [Holanda], núm. 25(2), p. 127-144. També disponible en línia a: <http://link.springer.com/article/10.1007/s10590-011-9090-0> [Consulta: 28 juny 2016].

Garrido, Juan María. (2012): «Glissando, un corpus anotat per a l'anàlisi de la prosòdia del català i del castellà. Descripció i primers resultats d'explotació». *Phonica* [en línia] [Barcelona: Universitat de Barcelona], núm. 8. <http://www.publicacions.ub.edu/revistes/phonica8/documentos/880.pdf> [Consulta: 28 juny 2016].

Gartner (2014). *Technology Overview for Text Analytics*. [en línia] <https://www.gartner.com/doc/2828817/technology-overview-text-analytics> [Consulta: 28 juny 2016].

Generalitat de Catalunya. (2015). *Més valor a la informació de salut de Catalunya (VISC+). Memòria Projecte* [en línia]. Barcelona: Generalitat de Catalunya. Departament de Salut. <http://aquas.gencat.cat/web/.content/minisite/aquas/projectes/antic_visc/memoria_visc_aquas2015.pdf> [Consulta: 28 juny 2016].

Generalitat de Catalunya (2016). *Dades massives per trobar un best-seller* [en línia]. Barcelona: Generalitat de Catalunya. Secretaria d'Universitats i Recerca. <http://universitatsirecerca.gencat.cat/ca/detalls/noticia/Dades-massives-per-trobar-un-best-seller> [Consulta: 28 juny 2016].

González-Aguirre, Aitor; Laparra, Egoitz; Rigau, German (2012). «Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base». A: *Proceedings of Sixth International Global WordNet Conference* [en línia]. Matsue, Japó. <http://www.slideshare.net/pathsproject/multilingual-central-repository-version-30-upgrading-a-very-large-lexical-knowledge-base> [Consulta: 28 juny 2016].

«Las redes sociales, medidores de la calidad de los fármacos». *El Correo.com* [en línia] [Bilbao] (17 març 2015). <http://www.elcorreo.com/bizkaia/tecnologia/investigacion/201503/17/redes-sociales-miden-calidad-20150317130117-rc.html> [Consulta: 28 juny 2016].

Moreno, Asunción [*et al*.] (2012). *The Catalan Language in the Digital Age = La llengua catalana a l'era digital* [en línia]. Heidelberg: Springer. (White Papers Series) <http://www.meta-net.eu/whitepapers/e-book/catalan.pdf> [Consulta: 28 juny 2016].

Padró, Lluís; Stanilovsky, Evgeny (2012). «FreeLing 3.0: Towards Wider Multilinguality». A: *Proceedings of LREC-2012* [en línia]. Istambul, Turquia. <http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf> [Consulta: 28 juny 2016].

Paniagua, Soraya (2013). «Un mundo de sensores, de los datos al 'Big Data'». *Telos: Revista de pensamiento sobre Comunicación, Tecnología y Sociedad* [Madrid: Fundación Telefónica], núm. 95 (juny-setembre), p. 94-96. També disponible en línia a: <https://telos.fundaciontelefonica.com/url-direct/pdf-generator?tipoContenido=articuloTelos&idContenido=2013062110130001&idioma=es> [Consulta: 29 juny 2016].

Rafel, Joaquim (1994). «Un corpus general de referència de la llengua catalana». *Caplletra: Revista Internacional de Filologia*. [València: Institut Interuniversitari de Filologia Valenciana], núm. 17, p. 219-250.

Reese, Samuel [*et al*.] (2010). «Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus». A: *Proceedings of LREC-2010* [en línia]. La Valleta, Malta. <http://www.cs.upc.edu/~nlp/papers/reese10.pdf> [Consulta: 29 juny 2016].

Sprenger, Timm O. [*et al*.] (2014). «Tweets and Trades: the Information Content of Stock Microblogs». *European Financial Management*, núm. 20, p. 926–957.

Taulé, Mariona; Martí, M. Antònia; Recasens, Marta (2008). «AnCora: Multilevel Annotated Corpora for Catalan and Spanish». A: *Proceedings of LREC-2009* [en línia]. Marràqueix, Marroc. <http://www.lrec-conf.org/proceedings/lrec2008/pdf/35_paper.pdf> [Consulta: 29 juny 2016].

Vázquez, Gloria; Fernández, Ana (2011). «Paralelización del corpus sensem: español-catalán». *Anuari de Filologia. Estudis de Lingüística* [en línia] [Barcelona: Universitat de Barcelona], núm. 1, p. 167-193. <http://revistes.ub.edu/index.php/AFEL/article/view/2252/2399> [Consulta 29 juny 2016].