Núria Castell Ariño
Usos socials del català
Barcelona: Institut d'Estudis Catalans, 2023, p. 67-73

DOI: 10.2436/15.7100.01.7
Data de recepció a la redacció:
20/02/2023

3. El tractament automàtic del llenguatge a casa nostra

Núria Castell Ariño[1]
Doctora en Informàtica
Professora titular jubilada
Universitat Politècnica de Catalunya

Actualment trobem arreu aplicacions basades en intel·ligència artificial (IA) i sembla que la IA sigui una tecnologia molt moderna. Però no és així. El concepte i el terme d'IA s'estableixen oficialment el 1956 a la coneguda conferència de Darmouth[2] (Veisdal, 2019). A iniciativa de John McCarthy, es van reunir experts de disciplines diverses (matemàtiques, enginyeria, física, psicologia, economia, sociologia) i van debatre durant un parell de mesos sobre com formalitzar l'aprenentatge i la intel·ligència humana per tal de construir màquines que poguessin simular el comportament humà.

Entre molts aspectes, es va plantejar ja en aquell moment de quina manera les màquines podrien utilitzar el llenguatge humà. Fins i tot es va pronosticar que en deu anys es disposaria de bons sistemes de traducció automàtica (TA). Aquest i altres pronòstics no es van complir pas i la IA, en general, i la TA, en particular, van caure en descrèdit a mitjans o finals de la dècada dels seixanta.

Per tant, el processament automàtic del llenguatge natural (PLN) ha estat un àmbit de recerca des de fa més de setanta anys. Durant molts anys, parlar de processament automàtic del llenguatge era sinònim de parlar de processament automàtic de l'anglès. Les primeres eines que van començar a funcionar raonablement bé eren analitzadors sintàctics de l'anglès. Eines que van servir de primers models per a construir analitzadors d'altres llengües.

La Comissió Europea va decidir apostar per la traducció automàtica per tal de facilitar la comunicació ràpida entre els estats membres i controlar el creixement de la plantilla de traductors humans. El 1978 es va posar en marxa un projecte preliminar amb un equip inicial que va preparar la proposta de projecte de

1. *A/e: nuria.castell@upc.edu*
2. <https://en.wikipedia.org/wiki/Dartmouth_workshop>

recerca col·laboratiu. Així el 1982 neix el projecte Eurotra[3] (Lau, 1987), centrat bàsicament en la lingüística computacional, amb l'objectiu de crear un sistema que permetés la traducció entre les diferents llengües oficials (9 en aquell moment). El castellà i el portuguès s'hi van incloure el 1986 i el projecte va estar finançat fins l'any 1992. El castellà va ser treballat per dos equips, un a Madrid i l'altre a Barcelona. El grup de Barcelona estava format per investigadors de diferents disciplines i de diverses universitats catalanes.

Per descomptat, el català mai no va ser inclòs en el projecte Eurotra, però el projecte va servir per a fer créixer els grups de lingüística computacional a casa nostra, donar importància a la recerca en processament del llenguatge natural i fer llavor de la recerca en processament automàtic del català que es va abordar des de la lingüística, la psicologia i la informàtica.

El projecte Eurotra va acabar sense la creació de l'esperat sistema de traducció automàtica. Un fracàs que la Comissió Europea va dissimular reorientant el finançament cap a projectes de generació de recursos necessaris per al processament automàtic del llenguatge. Així va néixer, entre d'altres, el projecte AcquiLEX (1989-92), seguit per AcquiLEX II (1992-95), projectes en què hi va haver participació de grups de recerca catalans.

Una lectura interessant sobre l'evolució de la traducció automàtica, projectes i grups de recerca durant el període 1985-2000 la trobem a Abaitua (2000).

Algunes grans empreses privades (IBM, Siemens, Fujitsu) també varen apostar per invertir en el desenvolupament de sistemes de traducció automàtica. En particular, Siemens va crear un equip de recerca a Barcelona per tal de desenvolupar el mòdul de castellà per al seu sistema METAL. Aquest grup es va escindir el 1992 per a formar l'empresa INCYTA, que ja va desenvolupar eines per al català i que actualment continua desenvolupant projectes lingüístics. Aquesta empresa va donar suport al primer diari que es va publicar en català utilitzant la traducció automàtica, el diari *Segre*. Aquest cas és força menys conegut que les publicacions traduïdes que van anar apareixent d'*El Periódico* i *La Vanguardia*. Aquests primers sistemes presentaven bastants errors de traducció, fins i tot després de passar per correctors humans. La majoria d'aquests errors avui en dia són anècdotes del passat, però no queda gaire llunyana alguna entrevista que m'han fet i que publicada en castellà feia referència a una tal «Nuria Castillo».

El finançament públic és fonamental per tal de fomentar la recerca bàsica atès que les empreses privades difícilment hi invertiran recursos si no hi veuen un benefici a curt o mitjà termini, per molt que tinguin capacitat econòmica

3. <https://cordis.europa.eu/programme/id/PRE_FWP_EUROTRA-1> <https://cordis.europa.eu/programme/id/FP2-EUROTRA-2>

suficient. En l'àmbit del processament automàtic del llenguatge a Catalunya val la pena mencionar l'empresa Intersoftware, creada per un visionari, Rafael Sala, que volia obtenir un sistema de lectura i indexació automàtica de notícies de diari. Des de 1977 i fins a 1985 (quan va fer fallida l'empresa) va finançar dos equips de recerca, un per a desenvolupar el lector òptic i l'altre per a desenvolupar el sistema de tractament de text que permetés indexar les notícies. El llenguatge a analitzar era el castellà, atès que la font de les notícies era el diari *El País*.

Com en el cas d'Eurotra, el projecte d'Intersoftware no va arribar a bon port, però va permetre la creació d'un equip interdisciplinari d'investigadors que en part es va integrar a la Universitat de Barcelona (UB) i a la Universitat Politècnica de Catalunya (UPC) quan l'empresa va tancar. Aquestes persones van seguir fent recerca a la universitat, participant en projectes europeus i nacionals i creant i/o ampliant els grups de recerca. La tesi doctoral, presentada a la UB, de Maria Antònia Martí Antonín (Martí, 1988) és fruit de l'experiència adquirida a Intersoftware. Cal destacar que es presentava un analitzador morfològic del català, una de les primeres, si no la primera, tesi en l'àrea de lingüística computacional que desenvolupava una eina per al català.

Conseqüència de l'experiència a Intersoftware són també dues tesis doctorals en l'àrea de la informàtica, les primeres en l'àmbit del processament automàtic del llenguatge natural a la UPC. La tesi d'Horacio Rodríguez Hontoria (Rodríguez, 1989) se centrava en el disseny d'un generador d'interfícies en llenguatge natural (el castellà). La tesi de l'autora (Castell, 1989) estava directament relacionada amb l'experiència prèvia: la comprensió automàtica de notícies de diari. Les notícies de treball eren del diari *El País* i, per tant, el llenguatge analitzat era el castellà. La tesi està redactada en català, una aposta per utilitzar aquesta llengua en publicacions tecnològiques malgrat tenir el hàndicap de fer-ne més difícil la difusió. L'ús cientificotecnològic del català i la valoració de publicacions en aquesta llengua per part de les agències avaluadores de la recerca és un tema a revisar si volem que el català sigui una llengua de comunicació real i no marginal.

Com ja s'ha mencionat anteriorment, el finançament públic és especialment necessari per a determinats àmbits de recerca. En el cas del processament automàtic del llenguatge natural això és especialment rellevant quan es tracta de llengües minoritàries com el català. A partir de mitjans dels anys noranta la Generalitat va anar creant els anomenats Centres de Referència en àmbits diversos amb l'objectiu de donar suport a la col·laboració entre els diferents grups de recerca de l'àmbit corresponent. En aquest marc, la Generalitat va crear el Centre de Referència d'Enginyeria Lingüística, que va aplegar els grups de recerca catalans tant de l'àrea del tractament del llenguatge escrit com de la parla. Les jornades celebrades l'any 2000 (Bozzo, 2001) van servir per a mostrar la recerca en curs de tots els grups. Malauradament, aquest centre va deixar de ser finançat i els

diferents grups van seguir la seva tasca de recerca amb projectes nacionals i europeus, i es van perdre en alguns casos les connexions que el CREL havia facilitat.

Ja no trobem més iniciatives de la Generalitat d'ampli abast orientades a potenciar el català en l'àmbit tecnològic fins al projecte AINA,[4] impulsat pel Departament de Polítiques Digitals de la Generalitat de Catalunya en col·laboració amb el Barcelona Supercomputing Centre (BSC). L'objectiu d'aquest projecte és generar recursos de veu per a entrenar algorismes basats en intel·ligència artificial que permetin desenvolupar sistemes com els assistents virtuals o traductors automàtics de parla en català.

Als inicis, el processament automàtic del llenguatge natural es basava principalment a incorporar coneixement lingüístic als sistemes informàtics. Era molt laboriós formalitzar aquest coneixement, tant més laboriós com més riquesa morfològica i sintàctica té la llengua. Per exemple, no eren necessaris analitzadors morfològics de l'anglès, però sí que es necessitaven per al castellà i el català. L'elaboració era manual, i hi col·laboraven lingüistes amb informàtics per a dissenyar i implementar el programari adequat.

El finançament de projectes dedicats a la creació de recursos lingüístics va permetre anar desenvolupant eines per al processament automàtic del llenguatge natural. La creació de lexicons a partir de diccionaris electrònics, els tesaurus, les xarxes semàntiques, els corpus textuals, els corpus paral·lels… van ser la base per a la creació d'analitzadors de text per a sistemes de comprensió del llenguatge. De manera similar es van anar creant recursos per al desenvolupament de sistemes de parla. Tenir més recursos, més dades, va anar acompanyat de millores en la capacitat de càlcul dels ordinadors i va possibilitar obtenir resultats en temps raonables.

Òbviament, el català sempre ha anat a remolc. No era fàcil trobar finançament per a investigar sobre una llengua minoritària i la recerca era de difícil publicació fins que no es va anar avançant en l'organització de trobades internacionals centrades en les llengües minoritàries. El processament automàtic del català ha anat avançant gràcies a la convicció dels equips de recerca locals que d'una manera o d'una altra han anat aconseguint finançament (per exemple, amb projectes del Ministerio que finançaven equips interuniversitaris per a investigar conjuntament el castellà, el català i l'eusquera, o fent treballs *paral·lels* aprofitant projectes europeus). Un resultat mostra d'aquestes estratègies va ser aconseguir desenvolupar un Wordnet,[5] la xarxa lexicosemàntica de referència durant molts

4. Projecte AINA: <https://www.projecteaina.cat/>
5. Base de dades Wordnet: <https://wordnet.princeton.edu/>

anys, per al català i connectar-lo amb altres llengües (castellà, eusquera i anglès) seguint l'enfocament multilingüe d'Eurowordnet.[6]

També algunes empreses han anat apostant pel català, més per convicció que pels possibles beneficis econòmics a curt termini. D'una banda, creant recursos (correctors, diccionaris…) i eines per a sistemes diversos (traducció automàtica, sistemes de diàleg…) i, de l'altra, traduint el programari al català. Algunes d'aquestes empreses han sorgit dels equips de recerca de les universitats on es generen recursos i eines que massa sovint troben moltes dificultats per a introduir-se al mercat.

Cal destacar iniciatives com Softcatalà,[7] associació sense ànim de lucre amb la missió de fomentar la presència i l'ús del català en tots els àmbits de les noves tecnologies (TIC). La seva filosofia es basa en el compromís amb el programari lliure i la llengua catalana.

Que grans empreses com Microsoft o Google incorporin versions en català del seu programari i recursos lingüístics només succeeix quan ho consideren rellevant, tenint en compte l'esforç que cal dedicar-hi. Però això també depèn de nosaltres, els usuaris, que hem de reclamar que volem treballar en la nostra llengua materna, que volem que les aplicacions tinguin la versió en català, que els cercadors prioritzin la llengua que ens interessa en lloc d'arraconar-la com s'ha posat de manifest que fan actualment (Cuesta, 2023).[8]

El processament automàtic del llenguatge natural ha anat evolucionant a mesura que s'han generat més recursos lingüístics i sobretot quan el volum de dades disponibles s'ha incrementat exponencialment i la capacitat de càlcul dels ordinadors permet obtenir resultats en temps curts. D'un enfocament purament lingüístic s'ha passat a la incorporació de tractaments estadístics, després a l'aplicació de tècniques d'aprenentatge (*machine learning*) i darrerament a l'ús de xarxes neuronals (*deep learning*). Aquests canvis han permès la millora substancial d'alguns sistemes com és el cas de la traducció automàtica.

Avui en dia disposem de sistemes automàtics de traducció (de text i de veu), xatbots, assistents virtuals de veu, resumidors de textos, correctors de text, classificadors de documents, detectors de spam en el correu electrònic, cercadors intelligents… Però no per a tots els idiomes ni amb la mateixa qualitat.

El català encara necessita incrementar la seva presència en el món de les noves tecnologies. El traductor de Google incorpora el català raonablement bé, però d'altres, com DeepL, no el tenen en compte. Alguns traductors són molt bons, però per a un nombre molt limitat de parelles de llengües.

6. Projecte EuroWordnet: <https://cordis.europa.eu/project/id/LE24003/es>
7. Associació Softcalà: <https://www.softcatala.org/>
8. Nota del curador: V. Partal (2023 [aquest volum]: 49-50).

La traducció automàtica té un problema, entre d'altres, que encara no ha resolt i que afecta el català. Quan una llengua té diversitat morfològica de gènere (cas del català) i l'altra no (cas de l'anglès), els estereotips i biaixos de gènere apareixen en els resultats de la traducció, la qual cosa reforça encara més els estereotips.

Els assistents virtuals de veu en català milloraran gràcies a projectes com AINA. En el cas dels xatbots comercials encara hi ha molta feina a fer, només cal intentar fer alguna gestió mitjançant el xatbot d'alguna empresa de serveis per a comprovar-ho. Aquests xatbots haurien de ser més bons tenint en compte que treballen en dominis restringits, en funció de l'àmbit de l'empresa, però les empreses no hi dediquen recursos per a millorar-los. Recentment s'ha fet públic un sistema que genera un fort debat, el ChatGPT[9] d'OpenAI, un sistema de diàleg basat en intel·ligència artificial que dona uns resultats força coherents en funció de la temàtica que se li planteja. Sorprenentment, incorpora el català, potser per la disponibilitat actual de moltes dades públiques en català com és el cas de la Viquipèdia, on tant esforç col·lectiu hem invertit i invertim (actualment és la vintena llengua en nombre d'articles).

Veient l'evolució del processament automàtic del llenguatge natural, on estem actualment i el cas particular del català, és evident que d'una banda som els usuaris de la tecnologia els que hem de reclamar la presència del català tant per a tenir les aplicacions en la nostra llengua com per a poder usar-la com una llengua més en les diferents aplicacions; però, de l'altra, els usuaris hem de ser proactius i utilitzar els recursos que ja estan disponibles (sistemes operatius, navegadors…) per a demostrar que són útils i necessaris. També és evident que per a fer avançar les llengües minoritàries necessitem la implicació de les institucions públiques i del Govern, no podem confiar només en el voluntarisme i l'activisme local de desenvolupadors i investigadors motivats, ni podem esperar que el sector privat destini molts recursos allà on no vegi un benefici tangible. Sumem esforços per a consolidar el català en l'àmbit tecnològic!

Referències bibliogràfiques

Abaitua, Joseba (2000). *Quince años de traducción automática en España* [en línia]: <https://paginaspersonales.deusto.es/abaitua/konzeptu/ta/ta15.htm>

Bozzo i Duran, Maria (2001). «Jornades del Centre de Referència d'Enginyeria Lingüística (CREL) (Barcelona, 4 i 5 d'abril de 2000)». A: *Estudis Romànics*, 2001, vol. 23, p. 377-378 [en línia]: <https://www.raco.cat/index.php/Estudis/article/view/8349>

9. ChatGPT: <https://openai.com/blog/chatgpt/>

CASTELL, Núria (1989). *Un model pel tractament de la informació temporal en un sistema de comprensió automàtica de notícies*. Tesi doctoral [en línia]: <https://upcommons.upc.edu/handle/2117/93237>

CUESTA, Albert (2023). «Per què els cercadors marginen el català a internet». Diari *Ara* [en línia]: <https://www.ara.cat/media/per-que-google-yahoo-duckduckgo-bing-cercadors-marginen-catala-internet_130_4599657.html>

LAU, Peter (1987). *EUROTRA, The machine translation project of the european economic communities* [en línia]: <https://vlex.es/vid/eurotra-machine-translation-project-216273433>

MARTÍ, Maria Antònia (1988). *Processament informàtic del llenguatge natural: un sistema d'anàlisi morfològica per ordinador*. Tesi doctoral [en línia]: <http://diposit.ub.edu/dspace/handle/2445/41667>

RODRÍGUEZ, Horacio (1989). *Guai: un generador automático de interfaces en lengua natural*. Tesi doctoral.

VEISDAL, Jørgen (2019). *The Birthplace of AI. The 1956 Dartmouth Workshop. Cantor's Paradise.* [en línia]: <https://www.cantorsparadise.com/the-birthplace-of-ai-9ab7d4e5fb00>