Autora Fina Freixa i Pujals

Anàlisi lingüística dels verificadors ortogràfics i dels diccionaris informatitzats en català

La presència del català en l'àmbit de la informàtica no ha deixat de créixer durant els darrers anys. El primer pas va ser la possibilitat de treballar amb tots els caràcters catalans (accents oberts, ce trencada, majúscules accentuades, ela geminada, etc.); més endavant, ja va ser possible utilitzar programes escrits íntegrament en català. Al principi, es podien trobar en català únicament programes fets a mida per i per a particulars, empreses o institucions dels Països Catalans. Actualment, però, podem constatar que el nombre de programaris i maquinaris disponibles en català augmenta de forma considerable amb la incorporació de productes d'àmplia difusió comercial que s'estan traduint a la nostra llengua.

Paral·lelament a aquest procés, s'ha començat a treballar des de diversos sectors en el tractament automatitzat de la llengua catalana. El català ha començat a ser present en l'àmbit de les indústries de la llengua.

En aquest context, s'ha de situar l'augment actual dels programes de verificació ortogràfica i dels diccionaris en suport informàtic. Per aquesta raó la Direcció General de Política Lingüística em va encarregar l'anàlisi lingüística d'aquests diccionaris i programes, anàlisi recollida en el document *Anàlisi lingüística dels verificadors ortogràfics i dels diccionaris informatitzats existents en català*, acabat a l'octubre de 1994.[1]

Objectius

Aquest treball tenia dos objectius bàsics. D'una banda, analitzar i, posteriorment, descriure, des d'un punt de vista lingüístic, els programes de verificació ortogràfica i els diccionaris informatitzats més representatius que hi ha disponibles al mercat.

De l'altra, posar a disposició dels serveis lingüístics un inventari dels productes existents acompanyat de la informació bàsica de cada producte (fabricant, característiques tècniques, característiques lingüístiques, etc.).

Verificador ortogràfic «versus» corrector ortogràfic

Una de les primeres coses que s'observen a l'hora de dur a terme l'anàlisi d'aquests productes és que les empreses no utilitzen el mateix terme per designar-los. Normalment, aquesta diversitat de denominacions (*diccionari ortogràfic, diccionari català, mòdul de llenguatge, corrector ortogràfic* i *verificador ortogràfic*) no respon a diferències qualitatives i funcionals reals d'aquests productes, ja que tots es basen en diccionaris o llistes de paraules i tenen unes característiques força semblants.

En el treball vam deixar de banda els termes *diccionari ortogràfic* i *diccionari català* que, tot i que reflecteixen l'eina bàsica del producte, la majoria de vegades no són prou adequats ja que no tenen en compte altres funcions que aquests programes realitzen (com ara partició sil·làbica, separació sil·làbica, detecció de

paraules repetides, llistes de paraules alternatives, notes i comentaris lingüístics d'ajuda, etc.). Vam deixar de banda també el terme *mòdul de llenguatge*, perquè forma part de la terminologia d'una empresa determinada, i ens vam centrar en les altres dues denominacions.

En termes generals, i sense entrar en precisions terminològiques, podem dir que un *corrector ortogràfic* és un programa que detecta, marca i corregeix els errors lingüístics, mentre que un *verificador ortogràfic* únicament detecta i marca aquests errors sense corregir-los.

Com es pot veure en el treball, a la pràctica, tots els productes existents actualment al mercat per al català es limiten a detectar i marcar els errors lingüístics. És cert que alguns proposen llistes de paraules alternatives i incorporen opcions de substitució automàtica o semiautomàtica de la forma errònia per la correcta, però sempre cal que l'usuari seleccioni prèviament quina és l'alternativa correcta. Per aquest motiu, i prescindint de les denominacions que els fabricants els puguin donar, en el treball parlàvem sempre de verificadors ortogràfics i no pas de correctors ortogràfics.

Tipus de verificadors ortogràfics

En informàtica, l'evolució tecnològica dels productes es compta per generacions. Cada generació inclou el conjunt d'ordinadors, de programes, etc. que correspon a una mateixa època i a una mateixa tecnologia.

En l'àmbit específic dels verificadors ortogràfics, podem distingir tres generacions diferents de productes. La primera generació de verificadors ortogràfics treballa a partir d'una llista de paraules o diccionari. Durant la revisió d'un text, el programa detecta i marca les paraules que no consten al seu diccionari. Per aquest motiu, també se'ls anomena *diccionaris ortogràfics* o *verificadors lèxics*.

La segona generació fa un pas endavant i incorpora mètodes heurístics per a la detecció d'errors morfològics i sintàctics, bàsicament errors de concordança molt simples i localitzats.

Finalment, la tercera generació de verificadors utilitza les noves tecnologies en intel·ligència artificial i es basa en sistemes experts per tal de dur a terme una anàlisi completa de les frases, amb la qual cosa es detecten molts més errors i es fan moltes menys deteccions errònies.

La qualitat de tots aquests productes depèn, evidentment, de la tecnologia que incorporen, però sobretot també depèn de la qualitat dels materials lingüístics utilitzats per elaborar-los (paraules, regles morfològiques i sintàctiques, etc.).

La majoria dels 23 productes analitzats en el treball es poden incloure, sense cap mena de dubte, en la primera generació de verificadors. Treballen comparant les paraules del text que revisen amb les que consten al seus diccionaris (alguns inclouen la possibilitat d'incorporar diccionaris d'usuari, temàtics, etc.) i marquen totes les paraules que no hi consten, tant si són correctes com incorrectes. Per tant, detecten errors dactilogràfics, ortogràfics i morfològics en la mesura que es tracta de paraules que no consten als diccionaris del programa. Poden incorporar altres tipus de revisions (per exemple, detecció de paraules repetides consecutives, de paraules que contenen números, etc.), solucions alternatives, comentaris i exemples d'ajuda, funcions addicionals, etc., però tots es basen en un o més diccionaris. Dins d'aquest gran grup, algun producte incorpora la detecció d'algun cas greu d'apostrofació incorrecta com ara «beveu'ne» i «sentir'ho». Tot i que es tracta de casos molt puntuals, vam creure convenient distingir-los de la resta i els vam tipificar com a productes de la primera generació, avançada.

En canvi, un grup reduït de productes incorpora funcions de detecció d'errors d'apostrofació, de contraccions, i dels casos més simples de combinacions de pronoms febles incorrectes. Aquests productes estan a cavall de la primera i de la segona generació ja que, d'una banda, incorporen funcions que van més enllà de la simple comparació de les paraules del text amb les dels diccionaris del programa, i de l'altra, no inclouen encara la detecció d'errors de concordança bàsics com ara la concordança del nom amb l'adjectiu, del subjecte amb el verb, etc. Amb tot, els vam tipificar com a productes de la segona generació, tot i que sabem que és discutible.

Metodologia

Per a l'elaboració d'aquest treball es va seguir una metodologia basada en les fases següents: una primera fase de preparació, durant la qual es va fer un esbós dels recursos tècnics i humans disponibles, es van elaborar els materials necessaris per a l'anàlisi, i es van fer els tràmits per a l'obtenció dels productes que havien de ser analitzats; una segona fase d'anàlisi pròpiament dita i, finalment, una tercera fase de redacció, revisió i presentació del treball.

Ens centrarem, tot seguit, en els aspectes més rellevants de la fase de preparació com són l'obtenció del corpus de treball, l'elaboració de la bateria de proves i el disseny de les fitxes que ens van permetre de recollir i presentar la informació.

Elaboració de la llista de productes

Una vegada fixats els objectius del treball, es va començar a treballar en l'elaboració del corpus de base. La llista de verificadors ortogràfics i diccionaris informatitzats existents en el mercat es va obtenir a partir del buidatge de documentació diversa com ara propaganda comercial, catàlegs de productes informàtics, llistes i relacions de maquinaris i programaris en català, bases de dades, etc.

Elaboració de la bateria de proves

Tot seguit, es va procedir a l'elaboració d'una bateria de proves que permetés d'analitzar els tipus d'errors lingüístics que detecten els programes, les alternatives que proposen, etc.

Es va confeccionar una primera bateria formada per tres tests diferents. El primer contenia prop d'un centenar de frases, una cinquantena de paraules complexes i una desena de sintagmes nominals. El segon contenia tres textos corresponents a varietats funcionals, i el tercer, set textos corresponents a varietats geogràfiques i generacionals diverses. L'objectiu del primer test era presentar tot tipus d'errors (dactilogràfics, ortogràfics, morfològics, sintàctics, semàntics, lexicogràfics, terminològics, etc.), en contextos diferents (situació dins la frase, caràcters diferents com ara versaleta, cursiva, etc.) per tal d'observar el funcionament dels programes. D'altra banda, els altres tests ens permetien de comprovar el comportament dels productes davant de la diversitat lingüística de la llengua catalana. En general, es van escollir fragments de textos molt breus per tal d'agilitar l'anàlisi. El segon test contenia fragments de textos administratius, jurídics i mèdics, la terminologia dels quals no és gaire especialitzada. Pel que fa al tercer test, recollia textos de varietats dialectals diverses, també molt breus, que contenien trets morfològics, lèxics, etc. característics d'aquestes varietats (per exemple, l'article salat, els possessius, els demostratius, les paraulotes, etc.).

Aquesta bateria inicial va ser modificada posteriorment durant la fase d'anàlisi dels productes per tal de millorar-la i d'adequar-la a les noves necessitats que

anaven sorgint. Pel que fa al primer test, s'hi va incorporar la versió correcta dels textos per tal de comprovar el funcionament del programa en cada cas; i s'hi van afegir altres tipus d'errors com ara paraules repetides consecutives, paraules que contenien nombres, paraules que contenien lletres en majúscules, etc. per tal de verificar altres funcions lingüístiques que oferien els productes analitzats. Finalment, s'hi va incorporar un quart test per a les proves d'ordenació alfabètica.

No es va crear cap test específic per a les proves de partició sil·làbica ja que es va preferir d'utilitzar els textos existents i forçar en cada cas particions sil·làbiques incorrectes.

Elaboració de les fitxes

Durant la realització del treball, es van dissenyar tres tipus de fitxes diferents que ens van permetre de recollir i presentar les informacions d'una manera clara, estructurada, concisa i homogènia: la fitxa tècnica, la fitxa de presentació i la fitxa resumida.

La fitxa tècnica és la fitxa utilitzada durant l'anàlisi dels productes. Ens va permetre de recollir el màxim d'informació sobre cada producte amb el mínim d'esforç i de temps. Conté els apartats i subapartats següents: dades generals (informacions comercials), característiques tècniques, característiques lingüístiques (tipus d'errors que detecta, diccionari, solucions alternatives, altres utilitats, observacions), revisió lingüística de les unitats que constitueixen el producte i dades sobre l'anàlisi. Cada epígraf conté un reguitzell d'especificacions on es poden consignar les informacions de cada producte. Durant l'anàlisi, es distingien les informacions provinents del buidatge de la documentació facilitada pel fabricant de les informacions provinents de l'anàlisi del producte.

La fitxa de presentació, que anomenàvem simplement fitxa, va ser el resultat de la simplificació de la fitxa anterior de cara a l'exposició dels resultats del treball. Es van suprimir de la fitxa anterior tots aquells aspectes que es van considerar innecessaris, excessivament detallats, massa tècnics o poc documentats a la majoria de productes. Aquesta fitxa, per tant, conservava l'estructura de la fitxa tècnica però sense l'apartat de *Dades sobre l'anàlisi*. Cal precisar, però, que una bona part de les dades tècniques dels productes, sobretot les que fan referència a les incompatibilitats, es van extreure de la documentació que acompanya els productes.

Finalment, la manca de recursos i de disponibilitat d'alguns dels productes per fer-ne l'anàlisi i la durada excessiva del projecte ens van fer replantejar la idea inicial d'analitzar des d'un punt de vista lingüístic tots els verificadors i diccionaris informatitzats existents per al català. Per això vam dissenyar un tercer tipus de fitxa que ens permetia consignar les informacions bàsiques dels productes que no es van poder analitzar. Aquestes informacions provenien íntegrament del buidatge de les fonts utilitzades per a l'elaboració de la llista de productes i de consultes telefòniques puntuals fetes a fabricants i distribuïdors. La fitxa resumida es va utilitzar també per als productes que són del mateix fabricant que algun dels productes analitzats, quan aquest ens va confirmar que efectivament es tractava del mateix verificador ortogràfic (cas dels programes d'IBM i de Lotus). No es van incloure en aquest apartat els productes que en aquell moment no es comercialitzaven (Printex, VOLC/DSSP i Diccionari català del Ditexto).

Forma de consulta

Es pot accedir a la informació consultant directament els apartats *3.1 Fitxes*

(de productes analitzats) i *3.2 Fitxes resumides* (de productes no analitzats), on hi ha les fitxes classificades per ordre alfabètic de productes.

Tanmateix, per facilitar la consulta a partir d'unes informacions determinades, es va incloure un quart apartat d'índexs on es poden trobar les informacions corresponents al nom del producte, al fabricant i a l'entorn informàtic del producte classificades segons el tipus de producte, el nom del producte, l'entorn i el fabricant.

Si qui consulta el treball té algun dubte sobre el significat dels termes tècnics utilitzats, pot consultar el glossari que hi ha al cinquè apartat, que està adreçat als usuaris amb coneixements elementals o nuls d'informàtica.

Nota

1. La Direcció General de Política Lingüística ha fet arribar aquest document als Serveis Centrals del Consorci per a la Normalització Lingüística, als 22 centres de normalització lingüística que en depenen, a les unitats de la Xarxa Tècnica de la Generalitat de Catalunya que atenen consultes, a les delegacions territorials del Departament de Cultura i als serveis lingüístics de l'àmbit socioeconòmic.

■