

Corrector tipogràfic, ortogràfic i gramatical de català

Resum del projecte

Versió 1.0

(<http://www.elcorrector.cat>)

GLiCom, Universitat Pompeu Fabra

(<http://glicom.upf.es>)

Resum: aquest document és una presentació general de la feina realitzada al llarg de la implementació del corrector ortogràfic i gramatical. Consta de tres parts informatives: una visió panoràmica del projecte, una descripció breu de l'arquitectura informàtica del programari, i una descripció també breu de les característiques i funcionalitats lingüístiques del corrector.

1 Taula de continguts

1	Taula de continguts.....	2
2	Visió panoràmica del projecte	3
2.1	Què és un corrector automàtic?.....	3
2.2	Què fa interessant El Corrector?	3
2.3	Característiques tècniques d'El Corrector	4
2.4	Característiques funcionals d'El Corrector.....	4
2.5	El Corrector en xifres.....	4
3	Breu descripció de l'arquitectura de processament	5
3.1	Mòduls de l'arquitectura i flux de processament.....	5
4	Funcionalitats lingüístiques del corrector.....	8
4.1	Definició lingüística d'El Corrector	8
4.2	Tipologia d'errors tractats	8
4.2.1	Exemples dels tipus d'errors tractats	9
5	Glossari.....	11

2 Visió panoràmica del projecte

El Corrector és resultat d'un concurs públic, convocat per la Generalitat de Catalunya (expedient núm.: SE/CTTI/51/05), per tal de realitzar el “desenvolupament i implantació d'un programari corrector ortogràfic i gramatical per a la llengua catalana que funcioni en diverses plataformes i doni servei a diferents aplicacions d'utilització per l'Administració”, segons el plec de clàusules del concurs. Aquest programari ha de ser segons el mateix plec de codi obert.

2.1 Què és un corrector automàtic?

El procés de correcció d'El Corrector es basa en el que anomenen processament del llenguatge natural o humà. Això vol dir que en fer la correcció s'intenta que l'ordinador simuli una petita part de les tasques que faria un corrector humà. Per a això, cal programar-lo perquè emuli mínimament la capacitat de comprensió lingüística dels humans. Amb aquest nivell mínim de comprensió, el programa podrà fer una sèrie de prediccions sobre allò que l'usuari o usuària pot voler dir, i, a partir d'aquesta comprensió, determinar si allò que vol dir està expressat en una forma acceptada per la normativa lingüística. Si no ho està, llavors ha de generar un missatge d'error i, idealment, una o diverses propostes de correcció.

Evidentment, un corrector automàtic no està exempt de problemes ni de mancances. La tecnologia actual permet modelar amb programes informàtics un part molt petita de totes les complexes operacions cognitives que fem els humans per comprendre el llenguatge. Tal com està avui la tecnologia, hauran de passar encara unes quantes dècades abans no siguem capaços d'acostar-nos una mica més a les habilitats dels correctors humans. Això, en canvi, no és obstacle per dir que un corrector automàtic com aquest permet millorar la qualitat ortogràfica i gramatical dels textos escrits en català.

2.2 Què fa interessant El Corrector?

L'arquitectura de processament d'El Corrector és modular tant pel que fa als diversos algorismes de processament com als recursos lingüístics que aquests entren (diccionaris i gramàtiques).

La manera com ha estat concebut aquest programari permet que:

- Tot i tenir 18 aplicacions i un servei web de correcció totes les versions d'El Corrector utilitzin el mateix motor de correcció
- Per modificar el comportament d'El Corrector des del punt de vista funcional, modificant un sol diccionari, una sola gramàtica o un sol algorisme es modifiquin totes les versions d'El Corrector (18 connectors més el servei web)
- Cada un dels mòduls de l'arquitectura pugui ser millorat independentment
- La creació de connectors per a nous entorns o aplicacions es pugui fer sense modificar gens el nucli del programari de correcció
- Un informàtic pugui treballar en algorismes i programari sense veure's afectat ni afectar els recursos lingüístics
- Un lingüista pugui millorar el comportament lingüístic d'El Corrector sense coneixements d'informàtica
- Sigui fàcil incorporar recursos lingüístics d'àmbits temàtics específics

2.3 Característiques tècniques d'El Corrector

Des del punt de vista tècnic, les característiques del corrector són les següents:

- El motor de correcció funciona en tres plataformes: Windows, Linux i Mac
- En totes tres plataformes existeix un editor bàsic per emprar El Corrector, per si no es disposa de cap de les aplicacions per a les quals hi ha connectors (vg. més avall)
- En Windows presenta connectors per a 10 aplicacions: quatre d'elles són de MS Office (Word, Excel, PowerPoint i Outlook); tres d'OpenOffice (Write, Spreadsheet i Impress); el Mozilla Thunderbird; el Mozilla FireFox; i l'editor de pàgines HTML i el navegador del SeaMonkey
- En Linux presenta connectors per a 7 aplicacions: tres d'OpenOffice (Write, Spreadsheet i Impress); el Mozilla Thunderbird; el Mozilla FireFox; l'editor de pàgines HTML i el navegador del SeaMonkey; i l'Emacs.
- En Mac presenta connectors per a l'editor de pàgines HTML i el navegador del SeaMonkey.
- Existeix un servei web amb arquitectura distribuïda que permet la correcció des de qualsevol ordinador que disposi d'un dels tres navegadors següents: Mozilla FireFox, MS Internet Explorer, SeaMonkey.

2.4 Característiques funcionals d'El Corrector

Des del punt de vista lingüístic, les característiques del corrector són les següents:

- El lèxic amb què treballa és de la segona edició del Diccionari de la Llengua Catalana de l'IEC. És, per tant, un corrector normatiu.
- Els criteris de correcció per a errors d'ús (lèxics o gramaticals) es basen en la proposta feta en el plec de clàusules del concurs i segueixen també criteris normatius. El nombre de tipus d'errors tractables és limitat. Els tipus d'errors tractats estan detallats al document anomenat *Especificacions d'errors per a El Corrector* (que podeu consultar a l'adreça web següent: <http://parles.upf.es/corrector/Downloads/AnnexA.pdf>)
- Permet activar i desactivar la detecció dels errors tipogràfics o dels gramaticals. No permet desactivar la correcció dels errors ortogràfics.
- Permet activar que la correcció es realitzi segons una de les quatre variants dialectals següents: balear, central, nord-occidental i valencià.

2.5 El Corrector en xifres

La taula següent mostra en xifres el volum de codi generat en el marc del corrector.

Llenguatge	Fitxers	Línies	Instr.	Com.	Docu.	Classes	Mètodes/Classe
C Sharp	411	138.197	45.247	8,1%	23,8%	559	7,72
Web (HTML, *.js, *.aspx)	174	28.693	5.389	5,3%	-	-	-
C++	6	3.633	1.453	7,9%	-	4	26,25
Java	25	3.909	2.193	20,2%	-	40	5,55
Altres (*.xul, *.js, *.css)	28	-	-	-	-	-	-

El diccionari conté més d'un milió de paraules i més de quaranta mil expressions multímot. La gramàtica de detecció conté més de 1300 regles.

3 Breu descripció de l'arquitectura de processament

L'arquitectura interna del motor de correcció d'El Corrector és completament modular. Exteriorment s'hi accedeix com si es tractés d'un únic motor, amb una sola API (Interfície de Programació d'Aplicacions), que és allò que permet fer invocacions dels serveis que ofereix el motor de correcció des d'altres programes.

El motor de correcció està format per una sèrie de mòduls connectats seqüencialment de manera que la sortida d'un és l'entrada del següent, tot i que la connexió no és directa ja que està controlada pel mòdul corrector principal (CotigMain). Cada un d'aquests mòduls té una funcionalitat diferent dirigida a resoldre una tasca més o menys concreta, i tots ells en conjunt realitzen el procés de detecció i correcció d'errors tipogràfics, ortogràfics i gramaticals.

Els diferents mòduls s'agrupen per proximitat i funcionalitat en més avall específiques, cosa que permet definir interfícies públiques individuals per a cada mòdul i també avaluar-los per separat. Això fa que cada un d'aquests mòduls, alhora, sigui fàcilment modificable o suprimible.

A més aquesta arquitectura permet que el responsable de cada mòdul tingui total llibertat sobre la implementació interna, l'únic requeriment és la implementació de la interfície corresponent, i la validació corresponent amb els tests funcionals d'integració.

Important: si esteu interessats en la modificació o ampliació d'algun dels mòduls o recursos (diccionaris o gramàtiques) d'El Corrector, heu de consultar també el document [Guia per al desenvolupament d'El Corrector](#).

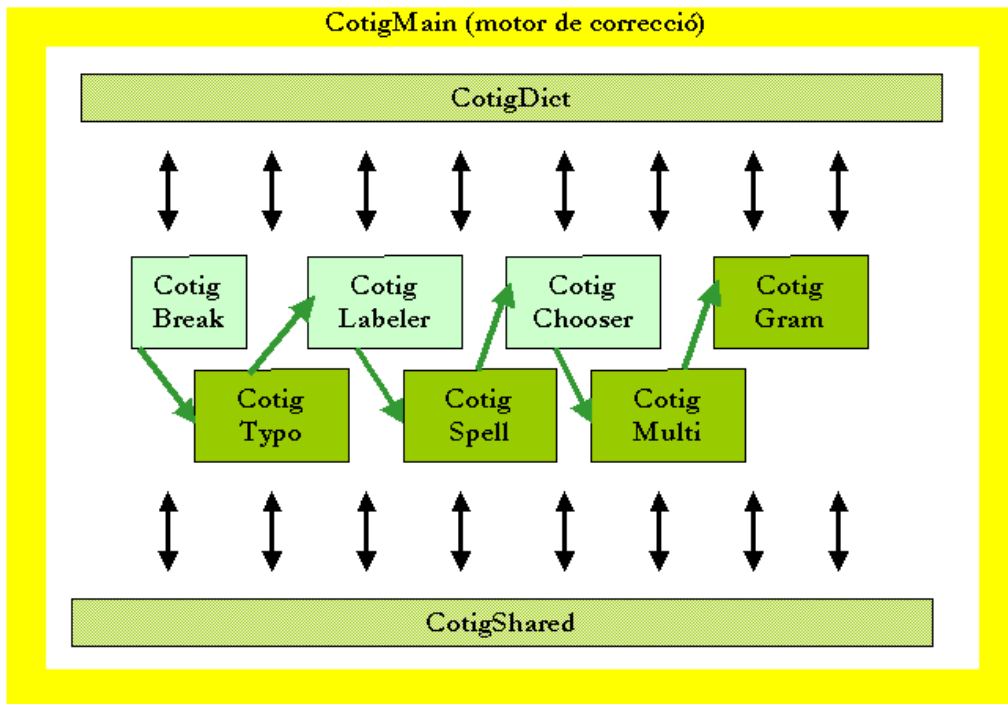
3.1 Mòduls de l'arquitectura i flux de processament

L'API del motor de correcció d'El Corrector està implementada en la biblioteca anomenada CotigMain. Qualsevol aplicació que vulgui comunicar-se amb el motor de correcció ho ha de fer a través d'aquesta biblioteca o a través d'alguna biblioteca que encapsuli aquesta biblioteca.

El CotigMain té la funció de gestionar el flux de processament de cada text. Determina com se segmenta el text en blocs de correcció (paràgrafs, frases, mots, etc.) i determina també quina mena de recursos es fan servir (diccionari, gramàtiques, etc.) en funció de les variants dialectals o els modes de correcció opcionals (tipogràfic i gramatical) triats per l'usuari/usuària.

Com dèiem inicialment, es tracta d'una arquitectura completament modular, de manera que qualsevol mòdul pot ésser fàcilment suprimit o substituït. Evidentment, aquests canvis implicaran un canvi en les prestacions d'El Corrector i han de ser realitzats per persones amb coneixements de programació i a partir de la documentació tècnica associada.

La figura que apareix en la pàgina següent reflecteix el flux de processament del text. A continuació hi ha unes taules que expliquen breument les funcionalitats de cada una de les llibreries.



La figura anterior mostra el processament seqüencial que el motor de correcció fa sobre els textos que es corregeixen. Els dos mòduls transversals al processament (CotigDict i CotigShared) són usats per totes les biblioteques de l'arquitectura. Aquests dos mòduls contenen informació lingüística o recursos informàtics necessaris per al funcionament del motor. La funcionalitat de cada mòdul es detalla breument en la taula següent.

Biblioteca	Funcionalitat
CotigBreak.dll	Segmentació de blocs, oracions i mots
CotigTypo.dll	Correcció tipogràfica (de caràcters)
CotigLabeler.dll	Etiquetatge de mots i elements textuais especials
CotigSpell.dll	Correcció ortogràfica
CotigChooser.dll	Desambiguació de lectura morfosintàctica
CotigMulti.dll	Tractament i correcció d'entitats multimot
CotigGram.dll	Correcció ortogràfica i gramatical contextuals

Pel que fa a les funcionalitats de les biblioteques comunes a tots els mòduls (CotigDict.dll i CotigShared.dll):

Biblioteca	Funcionalitat
CotigDict.dll	Gestiona la càrrega de diccionaris

CotigShared.dll	Defineix funcions comunes a totes les biblioteques del motor de correcció
------------------------	---

Per a més detalls sobre l'arquitectura podeu accedir al document anomenat *** i a la documentació tècnica associada a cada mòdul.

4 Funcionalitats lingüístiques del corrector

El Corrector és un corrector tipogràfic, ortogràfic i gramatical de caire general, és a dir, està pensat per ser emprat en la redacció de documents escrits en català sense que no pertanyin a una temàtica específica. Com tots els correctors, no està pensat per corregir qualsevol mena de text i com més es faci en el text un ús figuratiu de la llengua, més probabilitats tindrà l'usuari/usuària de no aconseguir-ne el resultat desitjat.

4.1 Definició lingüística d'El Corrector

El Corrector es defineix com a normatiu i diatòpic. És normatiu perquè segueix les normes lèxiques i gramaticals marcades per l'Institut d'Estudis Catalans i diatòpic perquè està pensat per corregir textos en les quatre grans variants dialectals del català. Aquestes variants són (per ordre alfabètic): balear, central, nord-occidental i valencià.

L'usuari/usuària d'El Corrector pot triar en quina variant dialectal vol escriure de manera que El Corrector tingui en compte quines formes farà servir. A la pràctica, aquesta adaptació a cada dialecte es reflecteix en el tractament de:

- formes verbals: per exemple, distinció entre *canto*, *cante* i *cant*
- aplicació d'algunes regles contextuais: per exemple, *per la tarda* és considerat un error en totes les variants tret de la valenciana

Altres variacions dialectals com poden ser les formes dels possessius (*meva/meua*, etc.), les dels articles (*el/es*, etc.), i les variacions lèxiques (*escombra/granera*, etc.) es donen per bones sigui quina sigui la variant triada per l'usuari/usuària.

4.2 Tipologia d'errors tractats

Els errors que tracta estaven definits inicialment en forma de llistat al *Plec de clàusules* que definia el concurs de licitació fet per la Generalitat de Catalunya. Aquest llistat s'ha traduït a unes especificacions concretes que detallen en quins contextos s'espera que El Corrector sigui capaç de detectar la presència d'un error, així com de marcar la paraula o paraules afectades i de fer-ne sempre que sigui possible una proposta de correcció adequada.

Els tipus de correcció que El Corrector implementa són:

- Correcció tipogràfica: detecció i proposta de correcció de l'ús incorrecte de determinats caràcters tipogràfics, com per exemple, l'ús del caràcter de l'accent tancat (´) en lloc del de l'apòstrof ('), o l'ús del punt (.) en lloc del punt volat (·) en la *ela* geminada.
- Correcció ortogràfica no contextual: detecció i proposta de correcció per a tots aquells errors que resulten en “no paraules”, és a dir, en seqüències de lletres que no formen part del català normatiu. Aquests errors s'anomenen no contextuais perquè el sol fet de ser presents en un text, independentment de les paraules que tinguin a dreta o esquerra, els delata.

Per exemple, si escrivim “ccasa” hem escrit una paraula que no forma part del català normatiu, i ho podem afirmar sense haver de mirar quines paraules l'acompanyen. Així mateix passaria si escrivíssim “sapigut”, que tot i ser habitual en algunes variants del català parlat, no és acceptada com a forma normativa per l'Institut d'Estudis Catalans.

- Correcció ortogràfica o gramatical contextual: detecció i proposta de correcció per a tots aquells errors que resulten en seqüències de paraules incorrectes (tot i

que les paraules per separat siguin paraules pròpies del català normatiu). Aquests errors s'anomenen contextuals perquè per determinar si la seqüència de paraules és correcta o no, cal fixar-se en les paraules que tenen a dreta o esquerra.

Per exemple, si escrivim “les raons econòmics de la decisió” només podem dir que “econòmics” és un error (o com a mínim que no encaixa en aquest context) si ens fixem que a la seva esquerra hi ha un article i un nom en femení plural.

4.2.1 Exemples dels tipus d'errors tractats

4.2.1.1 Error ortogràfic en general

Error que resulten de trabucar una o més lletres en escriure i que, com a resultat, donen lloc a una seqüència de lletres (“paraula”) que no pertany al català. Exemples: **paruala* (per *paraula*), **escenficació* (per *escenificació*), **froma* (per *forma*), etc. També es tracten errors en noms propis, gentilicis i antropònims: **Bracelona* (per *Barcelona*), **Jaon* (per *Joan*), etc.

4.2.1.2 Apostrofació d'articles

Error en la col·locació indeguda o la manca d'apòstrofs.

- Manca d'apòstrof: **el avi* (per ‘avi’ o “avi” *l'avi*), **la àvia* (per *l'àvia*), **en Oriol* (per *n'Oriol*), **na Anna* (per *n'Anna*).
- Apostrofació indeguda: **l'universitat* (per **la universitat*), **l'iode* (per *el iode*), **l'asimetria* (per *la asimetria*), **l'història* (per *la història*), **l'host* (per *la host*).

4.2.1.3 Apostrofació de preposicions

Error en la col·locació indeguda o la manca d'apòstrofs.

- Manca d'apòstrof: **de ahir* (per *d'ahir*), **de àvia* (per *d'àvia*), **de universitat* (per **d'universitat*), **de història* (per *d'història*).
- Apostrofació indeguda: **d'iode* (per *de iode*), **d'asimetria* (per *de asimetria*).

4.2.1.4 Apostrofació de pronoms febles

Error en la col·locació indeguda o la manca d'apòstrofs.

- Manca d'apòstrof: **no me agafa* (per *no m'agafa*), **li en pren* (per *li'n pren*), etc.
- Apostrofació indeguda: **s'em treu de sobre* (per *se'm treu de sobre*), **s'en ocupa* (per *se n'ocupa*), **dóna-m'en* (per *dóna-me'n*), **prenent-te ho* (per *prenent-t'ho*), **dóna-m'els* (per *dóna-me'ls*), etc.

4.2.1.5 Contracció de l'article masculí

Exemples:

- Manca de contracció: **la porta de el cotxe* (per *la porta del cotxe*), **la farina per a el pastís* (per *la farina per al pastís*), etc.
- Contracció incorrecta (davant de noms i adjectius començats per vocal): **el bastó del avi* (per *el bastó de l'avi*), **el remei pel hivern* (per *el remei per l'hivern*), etc.

Si voleu accedir a les especificacions detallades d'errors vegeu l'Annex A, document titulat *Especificacions d'errors per a El Corrector*. També el podeu trobar a l'adreça d'internet següent: <http://parles.upf.es/corrector/Downloads/AnnexA.pdf>.

5 Glossari

[NOTA: definicions extretes de la wikipèdia en català.]

API: Una **Interfície de Programació d'Aplicacions** (Application Programming Interface, API), és un conjunt de declaracions que defineix el contracte d'un component informàtic amb qui farà ús dels seus serveis.

Biblioteques DLL: **DLL** (*Dynamic Link Library*, Biblioteca d'Enllaç Dinàmic) és un format de fitxer de codi executable que és carregat a petició d'un programa per part del sistema operatiu. Esta denominació es referix als sistemes operatius Windows i és l'extensió amb què s'identifiquen els fitxers, encara que el concepte existix en pràcticament tots els sistemes operatius moderns.