

# Recursos d'ajut a l'edició

Ortografia, gramàtica i estil

Francesc Benavent Portabella  
Javier Gómez Guinovart  
Martí Quixal Martínez

PID\_00159191



Universitat Oberta  
de Catalunya

[www.uoc.edu](http://www.uoc.edu)



# Índex

<b>Introducció</b> .....	5
<b>Objectius</b> .....	6
<b>1. La revisió de textos i la verificació automàtica</b> .....	7
1.1. La revisió de textos .....	7
1.2. Origen, descripció i classificació dels errors .....	7
1.3. Verificació ortogràfica i gramatical, i verificació d'estil .....	10
1.4. Definició de la tasca de verificació automàtica .....	11
<b>2. La verificació automàtica de textos</b> .....	13
2.1. Tècniques de verificació automàtica .....	13
2.2. La verificació ortogràfica automàtica .....	15
2.2.1. Emmagatzemament del diccionari .....	15
2.2.2. Limitacions de la verificació ortogràfica automàtica .....	17
2.2.3. Correcció d'errors ortogràfics .....	17
2.3. La verificació gramatical automàtica .....	19
2.3.1. El reconeixement de patrons .....	20
2.3.2. Verificació gramatical de base estadística .....	23
2.4. La verificació estilística automàtica .....	23
2.4.1. Mètriques de llegibilitat .....	24
2.4.2. Lèxics controlats .....	25
<b>3. Disseny, implementació i avaluació d'un verificador automàtic</b> .....	27
3.1. Especificacions dels criteris de verificació .....	27
3.2. Disseny i desenvolupament de l'arquitectura .....	28
3.3. Avaluació del sistema de verificació .....	30
<b>Activitats</b> .....	33
<b>Exercicis d'autoavaluació</b> .....	33
<b>Solucionari</b> .....	34
<b>Glossari</b> .....	35
<b>Bibliografia</b> .....	37



## **Introducció**

La redacció de textos és una de les activitats més freqüents que fem amb l'ordinador. Gràcies als programes de processament de textos, l'ordinador esdevé una màquina d'escriure que conté eines d'ajut molt valuoses per a la redacció. Entre elles, hi trobem la col·locació automàtica del guionet a les paraules a final de línia, el còmput de paraules, el control dels canvis efectuats en diferents versions d'un mateix document, o, també, la verificació automàtica de llengua i estil. Dedicarem aquest mòdul a l'estudi de les eines de verificació ortogràfica, gramatical i d'estil automàtiques. La verificació automàtica es basa en un programari que fa servir tècniques de processament del llenguatge natural i té per objectiu ajudar a millorar la qualitat final del text.

## Objectius

Els objectius que assolireu amb els materials que componen aquest mòdul són els següents:

- 1.** Delimitar l'abast i els objectius de la verificació lingüística automàtica de l'ortografia, la gramàtica i l'estil, com a ajut a la redacció de textos.
- 2.** Analitzar les característiques dels errors d'ortografia, els gramaticals i els d'estil, especialment d'aquells més freqüents en l'escriptura amb ordinador.
- 3.** Examinar les tècniques emprades en la verificació automàtica de l'ortografia, gramàtica i estil en el marc general del processament de textos.
- 4.** Conèixer els aspectes essencials del procés de creació d'un verificador automàtic.

## 1. La revisió de textos i la verificació automàtica

Els verificadors lingüístics automàtics són un tipus de programari que serveix per comprovar l'adequació d'un text a la normativa o als criteris de redacció per als quals han estat dissenyats. Des del punt de vista de l'usuari o usuària final, la verificació consta de dues fases diferenciades: d'una banda, la identificació de les paraules del text susceptibles de contenir un error segons la normativa o de reflectir una discrepància o incoherència respecte dels criteris lingüístics o d'estil establerts; de l'altra, sempre que això sigui possible, la presentació d'una proposta de correcció.

### 1.1. La revisió de textos

La revisió de textos, feta per persones, s'entén com l'adequació d'un text "a les normes gramaticals de la llengua en què s'ha produït", i llavors es parla de **revisió gramatical**; i també com l'adequació d'aquest text "a la situació comunicativa de què forma part", i llavors es parla de **revisió d'estil** (Costa, pàg. 11). Segons Costa i altres (pàg. 64), la revisió de textos consta de diverses fases: lectura parcial, identificació d'errors freqüents i aspectes crítics, consultes a obres de referència, intervencions sobre el text i revisió general. Així mateix, és una activitat que pressuposa un domini de la llengua en totes les variants dialectals, rigor, objectivitat, coneixement de les obres de referència lingüística i dels recursos tipogràfics i convencions més esteses. També és convenient tenir un bon nivell de cultura i, no menys, prudència.

"Els termes *revisió* i *correcció* de textos són denominacions comunes per fer referència a una mateixa tasca".

J. Costa i altres (2006). *Curs de correcció de textos orals i escrits: pràctiques autocorrectives* (pàg. 5-6).

Evidentment, la verificació de textos automàtica és una activitat no gens comparable a la revisió de textos, però, en canvi, proporciona un ajut substancial a les persones que diàriament redactem textos que han de reunir un requisit formal mínim i no tenim l'oportunitat de fer-los passar per un procés de revisió manual fet per un tercer.

Seguint en part la distinció proposada per Costa i altres (2006), distingirem entre verificació ortogràfica i gramatical (revisió gramatical) i verificació d'estil (revisió d'estil).

### 1.2. Origen, descripció i classificació dels errors

En aquest subapartat examinarem les característiques principals dels errors ortogràfics, gramaticals i d'estil propis de l'escriptura amb ordinador.

Els errors o inconsistències que les persones fem quan escrivim un text amb ordinador poden ser motivats per un desconeixement de la norma o bé per una distracció. En el primer cas, en què la persona no sap com s'ha d'escriure correctament la paraula, expressió o signe gràfic en qüestió, es parla d'*errors de competència*; en el segon cas, en què la persona sap com s'ha d'escriure la paraula, expressió o signe gràfic en qüestió però, tot i així, s'erra, es parla d'*errors d'actuació*.

Els **errors de competència** tenen un vessant individual, ja que no tothom té el mateix nivell de coneixement de la normativa lingüística o dels criteris i convencions estilístics. Tanmateix, hi ha factors lingüístics que n'afavoreixen l'aparició, com ara el grau de disparitat entre ortografia i fonètica d'una paraula (*\*dons* per *doncs*), les discrepàncies entre normativa i parla (*\*coneixo* per *conec*), la interferència amb la normativa o el lèxic d'altres llengües (*\*tràfic* per *trànsit*, interferència de sentit amb el castellà *tráfico*, *\*el síndrome* per *la síndrome*), o la baixa freqüència d'ús d'una paraula.

Els **errors d'actuació** poden reflectir aspectes fonològics dels errors de la parla (*\*desmolaritzar* per *desmoralitzar*, amb intercanvi del tret fonològic de localització entre les consonants líquides), poden provenir d'una distracció visual (*\*problement* per *probablement*, amb el segment *blement* iniciat després la lletra *o* i no pas després de la lletra *a*), poden ser fruit d'un mal ús del teclat (*\*escriptutra* per *escriptura*, per haver premut alhora dues tecles pròximes en el teclat), d'una modificació en el text (si acabem fent una modificació en un part del text que requereix una modificació en una segona part del text però ens n'oblidem, per exemple, passem de *Aquest nen llegia un conte* a *\*Aquest nena llegia un conte* en lloc de *Aquesta nena llegia un conte*).

Independentment de les causes que provoquen que algú faci un error en la redacció d'un text, hi ha altres maneres de classificar-los. A continuació, exposem alguns criteris dels que es fan servir més freqüentment:

1) Errors que resulten o no en paraules: una mala redacció o mecanografiat poden fer-nos fer un error que resulti en una seqüència de lletres que no formi part de la llengua (*\*cassa* per *casa*), o en una seqüència de lletres que sí que en formi part (*#casa* per *caça*).

2) Tipus de transformació gràfica que ha patit la paraula o expressió afectada: se sol parlar de quatre transformacions bàsiques (també conegudes com a *transformacions de Damerau*). La inserció d'una o més lletres (*\*mediterrani* per *mediterrani*, *#té format quadrada* per *té forma quadrada*); la supressió d'una o més lletres (*\*exaurir* per *exhaurir*, *#mol interessant* per *molt interessant*); la substitució d'una lletra per una altra (*\*reb* per *rep*, *#pa de plat* per *pa de blat*); i la transposició de dues lletres adjacents (*\*perposició* per *preposició*). Aquestes transformacions també es poden aplicar a nivell de paraula sencera, no només de lletra o caràcter: supressió d'una paraula (*\*Li demano tingui comprensió* per *Li demano*

### Els errors

La revisió i la verificació de textos no impliquen necessàriament la identificació d'errors, també afecta les inconsistències formals o de contingut. Aquí, per simplificar, parlarem d'*errors* per referir-nos a tots dos, i farem la distinció quan sigui rellevant per a l'explicació.

### Nota

Al llarg de tot el document farem servir l'asterisc (\*) per indicar que una frase o expressió lingüística és incorrecta.

### Nota

Farem servir el símbol coixinet (#) per indicar frases, expressions o paraules que tot i ser correctes en termes estrictament formals no ho són en termes semàntics o pragmàtics.



que tingui comprensió), inserció d'una paraula (\**Pensa en que has de venir* per *Pensa que has de venir*) o substitució d'una paraula per una altra (\**No vindrà doncs està malalt* per *No vindrà perquè està malalt*).

3) Errors que afecten la frontera entre paraules: tenim l'error per encavalcament, que és la supressió de l'espai entre dues paraules (\**elnoi* per *el noi*), i l'error per partició, una separació indeguda mitjançant un espai (\**carr etera* per *carretera*).

4) La naturalesa lingüística de l'error: errors en la grafia (\**reduit* per *reduït*, \**barcelona* per *Barcelona*, \**trentadós* per *trenta-dos*), en la morfologia (\**volguer* per *voler*, \**lis diré* per *els diré*), en la sintaxi (\**abans de que vingui* per *abans que vingui*, \**el síndrome* per *la síndrome*), en la semàntica (#*tràfic* per *trànsit*) o en la pragmàtica (#*menos* per *menys* en un text formal, o #*nogensmenys* per *en canvi* dit en un diàleg de bar d'una telenovel·la).

5) Abast de l'error pel que fa a paraules: els errors (ortogràfics i gramaticals) poden diferenciar-se, entre els que poden ser detectats sense tenir-ne en compte el context (\**obvi* per *obvi*, \**coneixo* per *conec*) i els que només poden ser detectats tenint en compte el context (\**lis* per *els* és incorrecte, però *lis* és correcte a *flor de lis*, \**és* per *es* és incorrecte, però no a *aquesta és grossa*).

La classificació dels errors segons aquests o altres criteris tipològics serà convenient en funció de la utilitat que se li vulgui donar. Per exemple, la distinció entre el tipus de transformació no serà útil per explicar a l'usuari o usuària d'una aplicació final la causa de l'error, però en canvi serà útil per fer un estudi que ens permeti predir, a partir de tenir en compte els tipus de transformacions més freqüents, millors propostes de correcció. En canvi, la classificació per naturalesa lingüística de l'error ens pot ajudar a proporcionar un missatge d'error que acompanyi l'alarma disparada per l'eina i que faci comprendre el criteri lingüístic o d'estil pel qual es considera que hi ha un candidat a error.

Una distinció que, com veurem més endavant, és sovint rellevant per determinar la tècnica emprada per a la detecció de l'error a l'hora de dissenyar un programari per a la verificació automàtica de textos és la distinció entre errors que podem dir que ho són sense tenir en compte el context, i errors que només podem dir si ho són o no tenint en compte el context.

### Paraula

La definició de *paraula* en verificació automàtica sovint es redueix a una seqüència de lletres entre dos espais o un espai i un signe de puntuació, que és la definició més freqüent en processament del llenguatge natural.

### 1.3. Verificació ortogràfica i gramatical, i verificació d'estil

Entenem per *verificació ortogràfica i gramatical* la comprovació, i, si escau, la generació d'una proposta de correcció, de tots aquells errors relacionats amb l'incompliment de la normativa lingüística de la llengua en què s'ha produït el text. Així mateix, entenem per *verificació d'estil* la comprovació de tots aquells aspectes (errors o inconsistències) que incompleixen criteris i convencions relacionats amb el registre, la varietat dialectal o, per exemple, un estil corporatiu.

A continuació presentem una taula amb una llista d'exemples del que considerem errors (ERR) i inconsistències (INC) ortogràfics, gramaticals i d'estil. Noteu que no vol dir que aquests siguin la mena d'errors que tracten els verificadors automàtics.

#### Vegeu també

Els verificadors automàtics es tracten al subapartat 3.1.

Taula 1. Classificació dels problemes de redacció segons tipus i àmbit

Tipus	Àmbit	Exemple
ERR	ORTO	1) Deixem la <b>m´suica</b> barroca per sentir [...]. 2) T'agradarà l' <b>oba</b> que hem preparat [...]. 3) Ens fa molta <b>il.lusió</b> que passeu [...]. 4) El que vull assegurar <b>es</b> la possibilitat de recuperar dades [...].
ERR	GRAM	5) <b>Aniré</b> directe després de deixar la nena. 6) Va anar a recollir els <b>seves</b> germans a l'estació. 7) <b>L'hi</b> van dir que efectivament hi era. 8) Tu <b>contaves</b> que nosaltres vinguéssim, oi? 9) Hem trobat molt de <b>tràfic</b> a la N-340.
ERR	EST	10) Benvolguda Dra. Vives [...] <b>Apa! Fins aviat!</b> 11) Hem vist: <b>garses orenetes pinsans i xoriguers.</b>
INC	ORTO	12) <b>Van localitzar</b> el vaixell i <b>rescataren</b> els nàufrags.
INC	GRAM	13) Trob que portam sa <b>teva</b> sabata en es maleter.
INC	EST	14) Les " <b>meves</b> " són 'meves' i les 'teves' són 'meves' també.

No hi posem un asterisc per indicar que es tracta de frases amb errors o inconsistències perquè totes elles en tenen. Hi apareixen en negreta.

És important destacar que la frontera entre el que és un error i el que és una inconsistència no sempre serà clara i definida.

Així mateix, hem de tenir en compte que els criteris de classificació en errors o inconsistències que un podria establir seguint les recomanacions normatives o dels llibres d'estil no tenen per què coincidir amb la distinció error/advertiment necessària en el moment de la implementació d'un verificador automàtic. Com veurem més endavant, la precisió i la fiabilitat amb què es pugui modelar la predicció o simplement criteris funcionals idiosincràtics seran els que faran que determinats aspectes acabin convertint-se en missatges d'error o en advertiments.

#### 1.4. Definició de la tasca de verificació automàtica

La tasca de verificació automàtica se sol subdividir en tres fases. Per explicar-les ho farem a partir d'un exemple. Suposem que volem verificar automàticament la frase de l'exemple número 4 de la taula 1:

*\*El que vull assegurar es la possibilitat de recuperar dades.*

en lloc de:

*El que vull assegurar és la possibilitat de recuperar dades.*

El primer que cal és localitzar l'error. Per **localització de l'error** entenem l'acció d'identificar el punt en el text en què aquest comença i acaba. La localització servirà posteriorment per poder destacar visualment a l'usuari o usuària final l'error que es vol fer notar. Així en el nostre exemple l'error constaria d'una sola paraula (*es*) que comença al caràcter número 23 de la frase i acaba al número 24, aquí subratllats:

*\*El que vull assegurar es la possibilitat de recuperar dades.*

Cal tenir en compte que la localització de l'error no sempre coincidirà amb el context necessari per detectar-lo. Així, per detectar aquest error, ens caldrà comprovar si, donat el seu context lingüístic, la paraula *es* té més o menys probabilitats de tenir una lectura verbal (tercer persona singular del verb *ser*, per tant, *és*) o de tenir una lectura pronominal (pronom feble de tercera persona, per tant, *es*).

La segona fase del procés és la **classificació de l'error**, que consisteix a assignar un codi a la paraula o seqüència de paraules identificades com a errònies. Es tracta de categoritzar el problema detectat en una o més classes que permetin triar posteriorment el missatge que es presentarà finalment. En el nostre exemple, segurament podríem identificar l'error com a:

- *error ortogràfic,*
- *error d'accent diacrític,*
- *error per influència del castellà* (que no posa accent a la forma *es*).

Així n'hi podríem anar afegint tants com fossin possibles (tècnicament) i alhora rellevants per a l'aplicació final del programari.

El darrer pas en la verificació és la **generació de la proposta de correcció** que entenem com el procés mitjançant el qual el programari esdevé capaç d'oferir una alternativa de millora. En el nostre exemple, es tractaria de proposar la substitució de la forma *es* per *és*.

A l'hora de la veritat aquestes tres fases tampoc no són tasques tan independents com sembla. Així, la localització i la classificació de l'error sovint van lligades, ja que per poder detectar un error s'han de tenir en compte tant els elements lingüístics que el conformen (una o més lletres, una o més paraules, etc.) com els elements lingüístics que l'envolten, el context. Per exemple, per a la detecció d'errors com ara el de l'exemple 2 de la taula 1 (\*t'agradarà l'oba), només cal identificar que la seqüència de lletres *oba* no pertany al conjunt de paraules de la llengua catalana. Fet això, i sabent en quina zona del text es troba (en el nostre exemple és la quarta paraula de la frase) podem dir que es tracta d'un error ortogràfic (classificació) a la quarta paraula de la frase (localització).

Així mateix, per detectar un error com ara el de l'exemple 6 de la taula 1 (\**Va anar a recollir els seves germans*) haurem d'haver estat capaços de segmentar el text en paraules, assignar a cada una d'aquestes paraules la lectura més plausible en aquest context, i detectar que hi ha un error de concordança en el sintagma nominal format per aquestes tres paraules entre la paraula *seves* i les altres dues paraules que el componen. Això implica l'ús d'eines de processament del llenguatge natural.

Aquest exemple (el 6 de la taula 1) és interessant perquè ens permet veure també com la generació de propostes de correcció pot requerir sovint la reutilització d'informació lingüística derivada de la detecció i classificació de l'error. En aquest exemple, aprofitaríem els lemes de les paraules que conformen l'error, i les informacions morfològiques de categoria gramatical, gènere i nombre per poder generar propostes alternatives. Gràcies a una anàlisi que proporcioni la informació que presentem a la taula 2, podrem saber on comença i acaba l'error, i quina mena d'error és. El pas següent seria, doncs, generar alternatives a la seqüència \**els seves germans*, que, gràcies als nivells d'anàlisi lema, categoria gramatical, gènere i nombre podríem concretar en:

- *els seus germans*,
- *les seves germanes*.

Taula 2. Nivells d'anàlisi lingüística automàtica per a l'exemple \**a recollir els seves germans* (només els que són rellevants per a l'explicació)

Nivell descriptiu						
Identificador	...	3	4	5	6	7
Forma	...	a	recollir	els	seves	germans
Lema	...	a	recollir	el	seu	germà
Categoria gramatical	...	Prep	Verb	Art	AdjPoss	Nom
Gènere				Masc	Fem	Masc
Nombre				Plu	Plu	Plu

## 2. La verificació automàtica de textos

La verificació automàtica de textos és un procés informàtic en què el contingut lingüístic d'un document és revisat automàticament amb l'objectiu de detectar-hi errors i, si és possible, suggerir-ne la correcció.

Per tant, en la verificació automàtica podem diferenciar dos processos independents: d'un costat, la detecció de l'error (que equivaldria a les fases de localització i classificació explicades anteriorment) i, de l'altre, la generació d'una proposta de correcció.

En principi, aquesta descripció és vàlida independentment del nivell lingüístic al qual pertanyi l'error, però l'estat actual del tractament computacional del llenguatge fa que, quan és aplicada a **textos no restringits**, s'hagi de limitar la verificació automàtica als nivells tipogràfic, ortogràfic, gramatical (excloent sovint aspectes semàntics o pragmàtics) i estilístic.

Cal tenir en compte que, tot i tenir el mateix objectiu que la correcció de textos, la verificació automàtica es basa en una anàlisi superficial i, per tant, utilitza unes tècniques molt diferents de les estratègies (o destreses) que pot emprar un o una professional que en faci la revisió.

### 2.1. Tècniques de verificació automàtica

Des del punt de vista informàtic, l'anàlisi del text consisteix en l'exploració d'una seqüència d'elements de caràcter lingüístic, com són lletres, síl·labes, paraules o oracions. La unitat operativa concreta es pot triar segons el nivell lingüístic d'anàlisi, però el més habitual és que s'utilitzi la paraula. Les tècniques de verificació poden classificar-se en dos grups: tècniques no contextuais i tècniques contextuais.

1) **Tècniques no contextuais.** Es consideren tècniques no contextuais aquelles que per determinar la correcció o no d'un element lingüístic analitzen únicament les propietats d'aquest element, és a dir, les que es basen exclusivament en informació de caràcter intern. En aquest cas, la forma de l'element és determinant per decidir si és erroni o no. Per exemple, la forma de la paraula *gaxt* és suficient per determinar que es tracta d'un error, o més exactament, d'un element lingüístic que no forma part del conjunt de paraules de la llengua catalana.

2) **Tècniques contextuals.** Es consideren tècniques contextuals aquelles que determinen la correcció o no d'un element lingüístic a partir del context en què apareix, és a dir, a partir d'informació, que tot i ser local, és externa. En aquest cas, la forma de l'element és tan important com les propietats dels elements que l'envolten. Per exemple, el context de la paraula *nen* a l'oració *Els nen jugaven* és necessari (i suficient) per determinar que es tracta d'un error.

3) **Errors i advertiments.** Independentment del tipus de tècnica utilitzada en la detecció, contextual o no contextual, cal distingir entre allò que es mostrarà com a error i allò que es mostrarà com a advertiment. En el primer cas l'usuari o usuària rep la informació que aquella paraula és incorrecta i se li suggerirà una possible correcció. En el segon cas visualitzarà un advertiment que li farà saber que aquella paraula podria ser incorrecta o presentar una inconsistència i se li podrien explicar els contextos en què podria ser usada i aquells en què no.

El criteri per determinar si una estructura lingüística ha de mostrar-se com a error o com a advertiment és essencialment tecnològic. Si la tècnica utilitzada és capaç de determinar amb prou fiabilitat, sense intervenció humana, que es tracta d'una estructura incorrecta, es mostrarà com a error. Si la tècnica utilitzada no pot fer una predicció fiable de la intenció original de l'autor, o no pot desambiguar l'accepció usada sense preguntar a l'autor, s'indica com a advertiment i es deixa a criteri de l'autor o autora del text la possible correcció. Noteu que aquesta distinció no té necessàriament en compte la distinció que femiem al subapartat anterior entre errors i inconsistències.

Això no és massa diferent del que fan els correctors i correctores humans quan, en no poder determinar el significat que es volia transmetre en l'original o en no comprendre el sentit d'una determinada paraula o expressió, el marquen com a dubte (o haurien de fer-ho) per consultar-ho amb l'autor o autora del text.

Per exemple, tot i que la paraula *tràfic* és correcta si es refereix a l'intercanvi de mercaderies, és freqüent utilitzar-la erròniament per referir-se al flux de vehicles i, en aquest cas, caldria substituir-la per *trànsit*. La dificultat de determinar el sentit en què s'usa, fa recomanable indicar-ho com un advertiment que contingui l'explicació corresponent.

Per tant, en aquelles situacions en què les tècniques detectin un ús potencialment incorrecte, però no tinguin prou capacitat per reduir o eliminar els falsos positius, és aconsellable abstenir-se de tractar-ho com a error i en canvi fer-ho com a advertiment, de manera que només s'indiqui a l'usuari o usuària que seria convenient revisar aquella estructura.

En les seccions següents veurem exemples concrets de les tècniques que habitualment s'utilitzen per verificar errors ortogràfics, gramaticals i d'estil. Tot i que en cada un dels tipus de verificació posarem l'accent en alguna tècnica concreta, cal tenir en compte que no hi ha cap correspondència directa entre

un tipus d'error i un tipus de tècnica de detecció. La verificació d'un determinat error o de determinats tipus d'errors pot plantejar-se des d'angles diferents i amb objectius diferents.

Cal tenir en compte, però, que sovint s'associa correcció ortogràfica amb tècniques no contextuals, i correcció gramatical i d'estil amb tècniques contextuals. Això ve del fet que sovint s'associa la verificació ortogràfica a la verificació d'errors que resulten en no-paraules, fet que té una certa lògica si es té en compte que són els tipus d'error que era més fàcil tractar (i els que efectivament es tractaven) en els verificadors lingüístics "de primera generació".

## **2.2. La verificació ortogràfica automàtica**

Idealment, la verificació ortogràfica automàtica hauria de ser capaç de determinar si l'ortografia d'una determinada paraula és o no correcta. Per fer-ho només caldria comparar la forma que apareix al document amb la forma correcta de la paraula que previsiblement l'autor o autora tenia intenció d'escriure, amb el repte que això comporta.

Donada la impossibilitat de conèixer la intenció original, la verificació automàtica ortogràfica ha de plantejar-se amb uns objectius més modestos. Per exemple, determinar si la forma present al document es correspon amb la forma d'alguna de les paraules pertanyents a la llengua. Per això, les tècniques de verificació ortogràfica per al tractament d'errors que resulten en no-paraules acostumen a basar-se en la comparació del candidat a error amb una llista de paraules correctes emmagatzemada en forma de diccionari. Per a això són adequades les tècniques no contextuals, amb les quals únicament es té en compte la forma de la paraula sense analitzar el context en què apareix.

### **2.2.1. Emmagatzemament del diccionari**

La llista de paraules correctes que poden aparèixer en una llengua és molt extensa, habitualment se situa en l'ordre dels centenars de milers de formes, i fàcilment pot arribar a uns pocs milions si s'hi inclouen lèxics especialitzats i noms propis. Per això, la idea aparentment simple d'emmagatzemar totes les paraules correctes i identificar com a errors les paraules que no apareguin a la llista presenta algunes dificultats tècniques.

Bàsicament hi ha dues estratègies a l'hora de resoldre aquest problema: la utilització de **lemaris** i la utilització de **formaris**, que correspon a dues maneres diferents de codificar un **lexicó**. Les preferències cap a una o altra són determinades en cada moment per les característiques tècniques dels ordinadors utilitzats i per la importància assignada a l'eficiència a l'hora d'estalviar memòria o de maximitzar la velocitat d'accés al lexicó.

#### **1) Lemaris: diccionaris implícits**

En el moment de l'aparició dels ordinadors personals, la restricció tècnica més important dels sistemes de detecció d'errors ortogràfics amb tècniques no contextuais era la memòria: cap ordinador personal no podia emmagatzemar totes les formes flexionades ens els pocs megabytes de memòria disponibles que tenia. En aquests casos la solució idònia combina la utilització d'un diccionari d'arrels (lemes) i d'un conjunt de regles morfològiques que representen els diferents paradigmes flexius.

A l'hora de determinar si una paraula pertany a la llengua, aquests sistemes realitzen una anàlisi morfològica de cada paraula, comproven que l'arrel corresponent pugui ser flexionada, per obtenir la forma que volem comprovar, segons els paradigmes de flexió de la llengua, i en cas afirmatiu la paraula es dóna com a bona. En certa manera, aquesta solució basada en diccionaris de lemes conté implícitament la totalitat de les formes d'una llengua, i les genera dinàmicament quan necessita validar-ne una forma concreta.

Aquest sistema presenta com a avantatges la reducció en memòria del diccionari i una major facilitat a l'hora d'obtenir la llista de lemes correctes. En canvi, presenta com a inconvenients el risc de la sobregeneració, de manera que consideri com a correctes paraules que no ho són, i, sobretot, el cost computacional que suposa realitzar una anàlisi morfològica per a cada paraula, cosa que en limita l'aplicació en situacions en què el temps de resposta sigui crític.

## 2) Formaris: diccionaris explícits

A mesura que la potència dels ordinadors personals ha augmentat, amb la reducció consegüent del cost de la memòria RAM i l'augment fins a milers de megabytes de memòria, ha aparegut l'opció real d'utilitzar diccionaris de formes.

A partir d'una llista de formes flexionades possibles es crea el que es coneix com a **formari**. Aquest diccionari de formes inclou totes les paraules d'una llengua i permet determinar fàcilment si una paraula del document existeix o no en la llengua. En cas de no aparèixer a la llista es considera que es tracta d'un candidat a error i s'indica com a tal.

Fins i tot en els casos en què la memòria necessària no sigui un problema, cal garantir la velocitat d'accés, és a dir, el temps necessari per comprovar si una paraula és al diccionari. Si la llista s'hagués de recórrer seqüencialment ni els ordinadors personals més ràpids podrien revisar ortogràficament un text en pocs segons. Afortunadament hi ha tècniques d'indexació, com els **TRIE** (derivat de "reTRIEval"), que permeten reduir considerablement aquests temps. Per exemple, a partir d'un diccionari ordenat alfabèticament una cerca binària pot trobar qualsevol paraula 10.000 vegades més de pressa que una cerca lineal;



i mitjançant índexs de *hash* (l·listes associatives) es poden aconseguir temps d'accés constants independents de la llargada del diccionari, cosa que permet multiplicar per 100 la velocitat de la cerca binària.

Així doncs, tot i que els formaris requereixen més memòria i una feina prèvia considerable per a la preparació de recursos, el fet que puguin revisar centenars de milers de paraules per segon fa que siguin la solució més ràpida. Això els converteix en l'opció recomanable en entorns interactius d'edició de textos.

### 2.2.2. Limitacions de la verificació ortogràfica automàtica

Més enllà de les dificultats tècniques a l'hora d'emmagatzemar i consultar el diccionari, les principals limitacions d'aquest enfocament són les inherents a la simplificació de considerar que una paraula és ortogràficament incorrecta si no apareix en una llista tancada definida prèviament.

No és estrany que el document inclogui alguna paraula correcta que no hagi estat inclosa en el diccionari, produint-se un fals positiu durant la verificació, és a dir, la detecció d'un error inexistent. Els falsos positius sovintegen amb els noms propis i amb les paraules d'àmbits especialitzats. Per tal de reduir-ne la freqüència, els programes de verificació ortogràfica permeten l'ampliació personalitzada del diccionari; de manera que es puguin anar incorporant paraules al diccionari perquè en el futur no siguin considerades errònies.

D'altres vegades, pot fallar la identificació perquè l'error ortogràfic en una paraula n'origina una altra, diferent de la pretesa, però que per ser correcta es troba al diccionari (#*cinc* per *tinc*). En aquest cas el programa no detecta cap error ortogràfic i es produeix un fals negatiu. Si la seqüència vulnera les regles sintàctiques de la llengua (#*cinc gana* per *tinc gana*), l'error podrà ser identificat durant la verificació gramatical; en canvi, si no les vulnera (#*tinc grana* per *tinc gana*), el més probable és que passi desapercbut.

#### Vegeu també

La verificació gramatical es tracta al subapartat 2.3.

### 2.2.3. Correcció d'errors ortogràfics

Una vegada detectat l'error ortogràfic cal que el sistema automàtic proporcioni una o més propostes de correcció. En aquest cas es tracta que ofereixi una llista de paraules entre les quals es trobi la que es volia haver escrit. Per obtenir aquests candidats es parteix d'un model hipotètic sobre l'origen de l'error i s'inverteix el procés corresponent. En els errors ortogràfics que resulten en no-paraules s'assumeix el model clàssic que considera que la paraula original ha estat alterada mitjançant algun dels quatre processos de transformació descrits al subapartat 1.2.

Per obtenir la llista de paraules alternatives, només cal partir de la paraula desconeguda i, per a cada un dels caràcters, realitzar els processos inversos d'edició. La quantitat de combinacions possibles acostuma a ser molt elevada, però afortunadament la majoria d'elles poden descartar-se perquè generen

formes que no apareixen al diccionari. El resultat d'aquest procés és una llista de propostes de correcció que és mostrada a l'usuari o usuària final perquè en triï la correcta. Per exemple, els suggeriments de correcció per a la forma incorrecta *\*quarte* inclourien les paraules *quartet* (omissió de *t*), *quart* (inserció de *e*), *quarts* (substitució de *s* per *e*) i *quatre* (transposició de *t* i *r*). En alguns casos aquesta tècnica pot ampliar-se amb la incorporació de suggeriments per als errors que contenen més d'un d'aquests processos d'edició (*\*oetra* per *oberta*, amb ommissió de *b* i transposició de *t* i *r*) i per als que comencen per una lletra incorrecta (*\*pberta*).

A més, la majoria de sistemes no només obtenen una llista de correccions sinó que les ordenen segons la probabilitat de ser la correcció més adequada. Per fer això s'assumeix que no tots els processos d'edició ni tots els caràcters implicats són igual de freqüents. Així, sabem que els errors originats per la confusió entre *b* i *v*, entre *s*, *c* i *ç*, o d'ommissió de la lletra *h*, són més freqüents que la substitució d'una *a* per una *k*. Per formalitzar aquest coneixement els models incorporen models probabilístics de manera que els diferents processos d'edició tinguin diferents pesos. Aquests valors es poden assignar a partir de la similitud fonètica dels caràcters (per modelar errors estrictament ortogràfics) o de distància física entre les tecles (per modelar errors de mecanografiat). D'aquesta manera, a cada un dels candidats se li assigna un valor numèric corresponent a la suma dels pesos de les transformacions que han intervingut en la seva generació. Si els candidats s'ordenen segons aquest valor s'obté una llista encapçalada per les correccions més probables, és a dir, per aquelles formes obtingudes mitjançant els errors més habituals.

D'altres tècniques de correcció més sofisticades no es limiten a aquests quatre processos, sinó que tracten de descobrir les paraules correctes que més s'assemblen fonèticament o ortogràficament a l'error identificat. Per exemple, calculant el nombre de seqüències de dues o de tres lletres, ***n*-grames**, que tenen en comú dues paraules. Així, *\*pasetxar*, format pels trigramas [#*pa*, *pas*, *ase*, *set*, *etx*, *txa*, *xar*, *ar*#], en què el símbol # representa un caràcter d'inici i final de paraula, tindria tres trigramas en comú amb *passejar* ([#*pa*, *pas*, *ass*, *sse*, *sej*, *eja*, *jar*, *ar*#]) i només dos amb *pastera* ([#*pa*, *pas*, *ast*, *ste*, *ter*, *era*, *ra*#]). Igualment, es pot calcular la similitud fonètica entre dues paraules utilitzant un diccionari fonètic (una llista de paraules transcrites fonèticament) i un programa capaç de convertir la paraula incorrecta en la seva transcripció fonètica aproximada, i computant les seqüències de símbols fonètics que comparteixen.

Finalment, hi ha un tipus d'errors ortogràfics, normalment causats per les interferències d'altres llengües, que no es poden modelar a partir de simples processos d'edició, ja que la distància d'edició entre ells (v. distància de Levenshtein) és massa elevada i seria computacionalment massa costosa, o fins i tot impossible. Per exemple, davant de la paraula *\*vivenda* un sistema de verificació ortogràfica el podrà detectar com a error amb facilitat. Però a l'hora d'obtenir la seva proposta de correcció (*habitatge*) no el podrà obtenir automàticament a partir d'una o dues edicions. En aquests casos el més habitual és mantenir

una llista *ad hoc* d'errors freqüents i les correccions corresponents, que pugui ser consultada directament pel mòdul de detecció i generació de propostes de correcció.

### 2.3. La verificació gramatical automàtica

En la majoria d'eines de verificació automàtica es parla de *verificació gramatical automàtica* per fer referència a la detecció i generació de propostes de correcció dels errors que precisen tècniques de correcció contextuals. Aquesta distinció pot ser equívoca, ja que no tots els errors que precisen un tractament contextual són de naturalesa gramatical. Així, per exemple, per detectar si hi ha hagut un error en l'ús de la coma o de les cometes que obren i tanquen una expressió entre cometes, caldrà tenir en compte que abans o després del signe de puntuació en qüestió hi hagi o no un espai i una paraula. Per tant, per poder detectar l'error a:

*\*Vam veure la "seva" casa.*

en què sobra un espai entre la *a* de *seva* i les cometes de tancament, cal tenir en compte, un cop dividida la frase en elements textuais (també anomenats *tokens*), que a l'esquerra de les cometes de tancament hi ha la seqüència:

*COMETES OBERTURA + PARAULA + ESPAI*

i a la dreta

*ESPAI + PARAULA*

Aquí definim com a *verificació gramatical automàtica* el tractament de tots aquells aspectes que tinguin a veure amb l'ús correcte de la normativa i no tinguin a veure amb l'ortografia.

Així, doncs, considerarem errors com ara els inclosos en la fila relacionada amb errors gramaticals de la taula 1, que repetim aquí per comoditat.

#### Vegeu també

Les tècniques de correcció contextuals es tracten al subapartat 2.1.

Taula 3. Exemples d'errors tractables a la verificació gramatical automàtica

Tipus	Àmbit	Exemple
ERR	GRAM	<ol style="list-style-type: none"> <li>1) Aniré directe després de deixar la nena.</li> <li>2) Va anar a recollir els <b>seves</b> germans a l'estació.</li> <li>3) Va preguntar si efectivament aquella mare amb coratge homèric era a la sala. <b>L'hi</b> van dir que efectivament hi era.</li> <li>4) Tot i que t'oferies per fer la presentació, <b>contaves</b> que nosaltres vinguéssim [...].</li> <li>5) Hem trobat molt de <b>tràfic</b> a la N-340.</li> </ol>

Les tècniques emprades més habitualment en la verificació d'errors gramaticals, però també en la detecció de determinats errors ortogràfics, són les tècniques contextuals. Aquestes tècniques també s'anomenen *tècniques d'anàlisi sintàctica* (o *parsing*) i poden agrupar-se segons diversos criteris:

a) La profunditat o superficialitat de l'estructura amb què es fa l'anàlisi lingüística que sustenta el procés de verificació automàtica, i es parla de *tècniques d'anàlisi sintàctica profunda* o *superficial*.

b) L'abstracció del coneixement lingüístic mitjançant tècniques d'aprenentatge automàtic (de base estadística) o mitjançant la creació de regles per part d'experts, generalment lingüistes computacionals, segons un estudi de corpus o unes especificacions, i es parla de *tècniques de base estadística* o de *creació de gramàtiques amb regles manuals*.

Les tècniques que més es fan servir per la simplicitat computacional que presenten són les que pressuposen una anàlisi superficial, entre les quals es troben els **autòmats finits**. Aquestes tècniques permeten la detecció d'errors mitjançant reconeixement de patrons. Les tècniques basades en anàlisi sintàctica profunda són massa costoses computacionalment i no solen fer-se servir per a tasques com la verificació automàtica. El principal problema que presenten és que són difícils d'ajustar tant perquè no pateixin un bloqueig en el processament (per incapacitat de tractar determinades estructures) com perquè no acabin trobant correctes estructures que en realitat no ho són.

### 2.3.1. El reconeixement de patrons

La tècnica més utilitzada per al tractament automàtic de la verificació gramatical té un enfocament casuístic i es basa en el reconeixement de patrons. Això vol dir que el verificador recorrerà el text tractant de trobar les seqüències de paraules que segueixen unes determinades pautes d'error preestablertes.

En relació als errors de la taula 4, de primer veurem per què cal tenir en compte el context per detectar els errors dels tres primers exemples, i després veurem perquè pot ser opcional en els altres dos.

A les frases 1, 2 i 3 de la taula 3, la detecció dels errors implica considerar que:

a) A l'exemple 1 hi falta un complement direccional que a més es tracta d'un pronom feble que és part d'una forma verbal (*anar-hi*) lexicalitzada.

NOT *Pron HI* OR *CIRCUM\_LOC* + *Verb ANAR* + NOT *Pron HI* OR *CIRCUM\_LOC*

b) A l'exemple 2 cal tenir en compte que *seves* no concorda amb gènere ni amb *els* ni amb *germans* i que, si totes tres paraules han de formar un sintagma nominal, han de concordar en gènere i nombre.

*Det Masc Pl* + *Adj Fem Pl* + *Nom Masc Pl*

c) A l'exemple 3, per hipotetitzar un error sobre *l'hi* cal tenir en compte que està acompanyant un forma verbal (perifràstica) de *dir*, i que aquest verb ja té un complement directe (l'oració completiva *que efectivament hi era*).

*Pron EL* + *Pron HI* + *Verb DIR* + *QUE* + *Oració Subord*

Aquests patrons d'error poden utilitzar informació del nivell gràfic o fer ús d'alguna informació lingüística accessible per al verificador (lema, categoria gramatical, gènere o nombre, persona, temps, mode, etc.).

Els patrons d'error poden ampliar-se amb el corresponent suggeriment de correcció. Els patrons ampliat de la taula 4, en què el suggeriment de correcció s'indica a la dreta de la fletxa, serveixen per identificar i corregir quatre tipus d'errors gramaticals prou freqüents en català.

Taula 4. Patrons de correcció ampliat amb proposta de correcció

**Patrons de correcció ampliat amb proposta de correcció**

*a fi de que* → *a fi que*

*PENSAR en VINF* → *PENSAR a VINF*

*tant ADJ* → *tan ADJ*

*si ... PLUSCSUBJ ... PLUSCSUBJ* → *si ... PLUSCSUBJ ... CONDCOMP*

Hi trobem regles per detectar la inserció indeguda de la preposició de dins la locució *a fi que*, la introducció del complement infinitiu de *pensar* amb la preposició *en* (*\*Pensa en venir* per *Pensa a venir*), l'ús de *tant* davant d'adjectiu en lloc de *tan* (*\*No és tant fort* per *No és tan fort*), i l'ús del subjuntiu plusquamperfet

**Nota**

És cert que la frase *L'hi van dir que efectivament hi era* podria ser una frase correcta per referir a una frase com ara *A l'escola van dir a la Bruna que efectivament hi era*, en què *hi* estaria substituint el complement de lloc *a l'escola* però donem aquesta lectura per poc freqüent donada la major freqüència de confusió entre *li* i *l'hi*.

en lloc del condicional compost com a segona part d'una oració condicional amb temps compostos (\**Si ho haguessis dit ho hagués fet* per *Si ho haguessis dit ho hauria fet*).

Utilitzant diferents tipus d'abstraccions sobre les dades (com ara el lema *PEN-SAR* per a qualsevol forma flexiva del verb *pensar*, o *PLUSCSUBJ* per a qualsevol verb en plusquamperfet de subjuntiu) i símbols (com els punts suspensius per a qualsevol seqüència de paraules), els patrons assoleixen un grau de generalització que els permet de detectar i corregir un bon nombre de seqüències incorrectes.

D'altra banda, el tractament dels errors de les frases 4 i 5 de la taula 3 podria fer-se mitjançant tècniques no contextuals assumint que:

- Tractar-los tenint en compte el context requeria segurament l'ús de tècniques d'anàlisi semàntica o pragmàtica, les quals són encara poc precises i robustes per ser emprades en aplicacions reals.
- És fàcil confondre *comptar* amb *contar* i *trànsit* amb *tràfic*, per proximitat fonètica i per interferència amb el castellà.
- Com que es tractaria d'errors amb un grau de precisió baix, perquè no tenim en compte el context per detectar-los, podríem tractar-los com a advertiments i no pas com a candidats a error.

Cal tenir en compte que sempre podríem complicar les regles de detecció per a aquests dos errors si per exemple requeríssim que el verb que segueix *contar* estigui en subjuntiu perquè saltés la regla:

*Verb CONTAR + QUE + Verb Subj* → *CONTAR per COMPTAR*

Aquesta regla es podria sustentar en un estudi de corpus, per exemple, que mostrés que l'estructura *Verb + QUE + Verb Subj* és més freqüent amb *comptar* que no pas amb *contar*. Evidentment, de seguida ens podrien sortir contraexemples, però aquest és el punt en què s'ha de valorar i decidir si un prefereix cobertura (detectar un major nombre d'instàncies de l'error) o precisió (detectar-ne menys però fer-ho millor).

En qualsevol cas, perquè el verificador sigui capaç de detectar un error, aquest ha de correspondre a un dels patrons prèviament especificats; és a dir, perquè la verificació sigui efectiva i factible, cal anticipar la mena d'errors que poden aparèixer, i no tots els errors gramaticals són fàcils de preveure.

### 2.3.2. Verificació gramatical de base estadística

Les tècniques de detecció d'errors gramaticals de base estadística es fonamenten en el reconeixement de patrons, però en lloc de definir aquests patrons de manera manual, es fa a través d'abstreure coneixement estadístic derivat de corpus. Així, un algorisme habitual és caracteritzar en termes d'*n*-grames les estructures morfosintàctiques més freqüents d'una llengua, és a dir, modelar la llengua correcta, i després cercar en un text que s'ha de corregir tots aquells *n*-grames que no formen part del model de llengua correcta.

#### Lectures complementàries

Per aprofundir en les tècniques de detecció de base estadística, podeu consultar:

M. Gamon i altres (2009, juny). "Using Statistical Techniques and Web Search to Correct ESL Errors". *CALICO Journal* (vol. 26, núm. 3).

C. Leacock i M. Chodorow (2003). "Automated grammatical error detection". A: M. D. Shermis, J. Burstein (ed.). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Així, si desenvolupéssim un algorisme com aquest per al català, segurament trobaríem que la paraula *minuts* apareix molt més sovint després d'una xifra o d'un cardinal que no pas la paraula *menuts*. A través d'aquest coneixement estadístic, amb el programari adequat, podríem arribar a detectar en un text un error com ara

*#Va trigar més de 15 menuts a arribar.*

i proposar la correcció de *menuts* per *minuts*.

### 2.4. La verificació estilística automàtica

Un altre aspecte en què la verificació automàtica pot suposar un complement a l'edició és mitjançant l'ajut de la revisió d'estil. La revisió d'estil permet controlar tant el nivell de llegibilitat d'un text, com el compliment de determinades convencions de redacció definides per un diari o una editorial.

Des d'un punt de vista més abstracte, cal entendre l'estil com un conjunt de característiques objectives del document, tant quantitatives com qualitatives, associades a un determinat registre lingüístic o varietat estilística. El control d'estil permet associar al document un determinat estil predefinit i mesurar les discrepàncies que presenta en relació a altres documents prototípics. D'aquesta manera podem introduir els canvis necessaris per apropar-lo a l'estil desitjat.

Abans que el programa pugui realitzar la verificació estilística del document, cal especificar la categoria a la qual ha de pertànyer el text. A partir d'aquest moment el programa porta a terme la verificació comparant les característiques del document en qüestió amb els trets lingüístics que el sistema considera preceptius per a la categoria seleccionada.

En general, la categoria es pot seleccionar a partir d'un nombre de models estilístics predefinitos del programa. Però en alguns casos, el verificador permet crear nous models estilístics, assignant els valors desitjats a les característiques lingüístiques utilitzades per definir els diferents estils. Els trets utilitzats poden incloure mètriques com ara el nombre màxim de paraules per oració, la presència o absència d'expressions col·loquials, o el nombre màxim de sintagmes preposicionals consecutius; trets que intenten modelar i caracteritzar diferents variants estilístiques.

### **2.4.1. Mètriques de llegibilitat**

Entre les primeres tècniques utilitzades per a la verificació d'estil automàtica es troben les mètriques de llegibilitat, valors numèrics que permeten obtenir el grau de dificultat de lectura del sentit d'un text. Aquest grau de dificultat es mesura mitjançant un valor numèric obtingut a partir de la combinació de certs factors lingüístics quantificables, com la longitud de les oracions, la longitud de les paraules, la quantitat de preposicions en una frase o la raresa (poca freqüència d'ús) de les paraules.

Aquestes mètriques són el resultat d'estudis de llegibilitat curosos basats en la detecció de regularitats estadístiques que mostren els textos respecte de diferents trets obtinguts del text. Un dels seus objectius consisteix a elaborar fórmules de llegibilitat, contrastades empíricament amb mètodes estadístics, que serveixen per predir paràmetres estimatius objectius del grau de dificultat d'un text en funció del seu estil.

La fórmula de llegibilitat més coneguda és la proposada per Rudolph Flesh per a l'anglès. Aquesta fórmula determina l'índex de llegibilitat (IL) d'un text a partir de la mitjana de síl·labes per paraula (SP) i de la mitjana de paraules per oració (PO):

$$IL_{\text{Flesh}} = 206,835 - (84,6 \times SP) - (1,015 \times PO)$$

Aquest índex genera un valor entre 0 i 100 correlacionat amb la facilitat de lectura, de manera que els valors més baixos s'obtenen amb els textos més complexos:



Taula 5. Relació entre els índex de llegibilitat i els graus de dificultat

Llegibilitat	Dificultat	Nivell equivalent
90-100	molt fàcil	estudiant d'11 anys
80-90	fàcil	
70-80	més aviat fàcil	
60-70	estàndard	estudiant de 13 a 15 anys
50-60	més aviat difícil	
30-50	difícil	
0-30	molt difícil	graduat universitari

Hi ha una variant coneguda com a *Flesch-Kincaid grade level*, que retorna un valor directament comparable amb la numeració utilitzada en el sistema educatiu americà, des del *1st grade* fins al *12th grade*.

L'índex Gunning-Fog utilitza tant la mitjana de paraules per oració (PO) com la proporció de paraules complexes (PC). Definint com a complexes aquelles paraules de tres o més síl·labes, sense comptar noms propis o compostos:

$$IL_{\text{Gunning-Fog}} = 0,4 \times (PO + 100 \times PC)$$

L'índex Coleman-Liau que utilitza la mitjana de paraules per oració (PO) i la llargada mitjana en caràcters per paraula (CP):

$$IL_{\text{Coleman-Liau}} = 5,89 \times (CP - (29,5/PO))$$

O, l'*automated readability index* (ARI) que també utilitza la mitjana de paraules per oració (PO) i la llargada mitja en caràcters per paraula (CP):

$$IL_{\text{ARI}} = (4,71 \times CP) + (0,5 \times PO) - 21,43$$

Els resultats de tots aquests índexs acostumen a ser similars i, tot i ser prou fiables i objectius, només cal entendre'ls com uns indicadors orientatius per conèixer la complexitat formal del text analitzat.

#### 2.4.2. Lèxics controlats

Una altra tècnica important en la verificació d'estil automàtica és la utilització de lèxics i estructures controlades. La idea és utilitzar les mateixes tècniques utilitzades en la verificació ortogràfica i gramatical, però per validar que el vocabulari i les estructures utilitzades siguin les corresponents al registre estilístic seleccionat.

Un exemple d'aplicació es verificar l'ús d'una determinada varietat dialectal. En aquest cas, el diccionari ortogràfic hauria d'incloure per a cada paraula informació relativa a les variants en què és considerada pròpia (per exemple, *granera* seria vàlida en català balear i valencià, i *escombra* en català oriental i nord-occidental). En el cas que el text inclogui alguna paraula, que, tot i ser present al diccionari, no sigui vàlida de la varietat dialectal seleccionada, l'ordinador la marcarà com a incorrecta i, en alguns casos, hi suggerirà una alternativa. El diccionari pot incloure, a més de varietats dialectals, els registres en què pot aparèixer cada una d'elles. Això permet a l'usuari o usuària indicar que el text que està escrivint és un document formal, col·loquial, familiar o tècnic, i fer-ne la verificació segons aquests criteris.

En certa manera la idea és etiquetar els recursos lèxics de manera que el verificador ortogràfic utilitzi únicament un subconjunt del diccionari, el corresponent al registre o variant a què pertany el document. El mateix concepte pot aplicar-se al verificador gramatical: en aquest cas, els recursos són els patrons sintàctics o les regles gramaticals, que organitzades per criteris d'estil, s'apliquen únicament als textos d'un determinat registre.

### 3. Disseny, implementació i avaluació d'un verificador automàtic

Per acabar d'integrar els conceptes vistos fins ara i obtenir una visió general de què és un *verificador automàtic*, presentarem succintament el procés de disseny i implementació d'*El Corrector*, un corrector ortotipogràfic, ortogràfic i gramatical normatiu.

En aquest apartat presentem i expliquem les tres fases essencials de què consta el procés de creació d'un verificador automàtic: en primer lloc, la definició d'especificacions; en segon lloc, el disseny tècnic i el desenvolupament del motor de processament, i, finalment, el procediment d'avaluació.

#### El Corrector

Fou desenvolupat pel Grup de Lingüística Computacional de la Universitat Pompeu Fabra, amb la col·laboració de Barcelona Media Centre d'Innovació; en gran part *El Corrector* és el resultat del finançament rebut a partir d'un concurs públic convocat per la Generalitat de Catalunya (SE/CTTI/51/05).

#### 3.1. Especificacions dels criteris de verificació

Les especificacions d'un programa informàtic són essencialment una descripció del que s'espera que faci aquest programari. Així, doncs, les especificacions d'implementació d'un verificador automàtic seran essencialment de caràcter lingüístic, però també algunes que tindran en compte la mena d'interacció que esperem que l'usuari o usuària final pugui tenir amb el programa, i les que tindran en compte l'entorn de treball en el qual es farà servir el verificador. Aquestes darreres tenen a veure amb aspectes com ara el sistema operatiu en el qual haurà de funcionar, l'eina d'ofimàtica en la qual s'integrarà, o saber si ha de funcionar o no en una instal·lació en xarxa. D'aquests aspectes més generals i aplicables al desenvolupament de qualsevol eina de verificació, no en parlarem aquí.

A l'hora de definir les especificacions d'*El Corrector* es va partir de la descripció que en va fer la Generalitat de Catalunya en el document que feia públic el concurs de licitació. Com a criteris generals el concurs establia una sèrie de requeriments tècnics i de llicència, com ara que fos de codi lliure i obert, modular (separant funcionalitats i interfícies gràfiques), multiplataforma (Linux, Mac i Windows), etc., i una sèrie de requeriments sobre la funcionalitat lingüística: que fos d'ús general, que tingués en compte les formes adequades a la llengua l'estàndard de les varietats dialectals del català (finalment es van considerar les variants balear, central, nord-occidental i valenciana), que fos normatiu, que corregís errors ortotipogràfics, ortogràfics i gramaticals, que permetés un mínim d'interacció amb l'usuari o usuària final (activació i desactivació de la correcció gramatical, bloqueig de l'activació de determinats avisos, correcció interactiva, etc.).

Pel que fa a l'especificació dels errors ortogràfics i gramaticals que s'exigia que es tractessin, el document del concurs de licitació proporcionava una llista d'errors el tractament dels quals havia d'estar garantit i una llista d'errors el

tractament dels quals era d'interès. En la primera llista, s'hi incloïen errors com ara la correcció ortogràfica del lèxic comú, topònims, antropònims, etc., apostrofació d'articles, preposicions i pronoms febles (*\*el home, \*de ahir, \*m'el pren*), contracció de l'article masculí, guionets en els numerals, morfologia de noms, adjectius i verbs (*\*altruïste, \*reflexava, \*recullir, \*coneguent*, etc.), ortografia dels pronoms febles en general, perífrasi d'obligació (*\*tenir que, \*haver-hi que + infinitiu*), omissió de la conjunció *que* (*\*Us agrairé estúdieu l'informe*). La segona llista incloïa errors com ara l'ús de diacrítics (*és/es, dones/dónes*), l'ús de l'article neutre (*\*amb lo útil que és*), canvi i caiguda de preposicions (*\*Les hem acostumades a que mengin soles*), ús pronominal inadequat (*\*S'ha caigut del cavall*), etc.

El que ens interessa destacar aquí és que la realització d'aquestes llistes es va fer a partir de l'experiència que tenien les persones que van redactar les bases del concurs. Es tracta de les preguntes i errors més freqüents atribuïts a les persones que es dirigeixen al Servei de Consultes Lingüístiques de la Generalitat de Catalunya. Una altra manera de fer aquestes llistes, però, hagués estat seguint un procés de compilació, anotació i anàlisi d'un volum de textos considerable redactats per parlants identificables amb l'usuari o usuària final d'aquesta eina. Com us podeu imaginar, aquest procés és molt costós i lent i, de fet, és un tema pendent de resoldre en àrees de treball i recerca com ara l'estudi dels trets formals de la llengua en les diverses etapes d'adquisició del llenguatge, tant aprenents de segones llengües com de primeres llengües.

Un dels rols fonamentals de les especificacions és la creació d'una bateria de proves que serveixi per garantir la qualitat i l'adequació funcionals del programari de verificació a cada nova versió, i també durant el procés de desenvolupament. Així, per exemple, podem tenir conjunts de textos que ens permetin determinar els percentatges de cobertura i precisió del programari donades una sèrie de frases que sabem que exemplifiquen els errors que volem tractar.

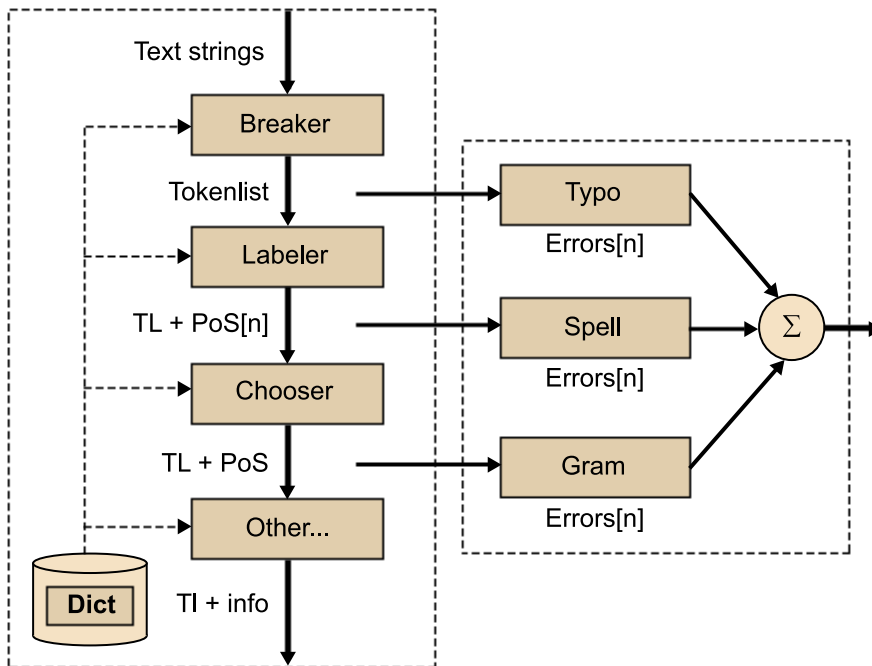
Cal tenir en compte que idealment aquestes llistes no han de tenir només els errors que el programari tracta correctament, sinó tots aquells que pensem que segons les especificacions hauria de ser capaç de tractar. Tampoc no haurien de contenir exemples d'errors que ja d'antuvi hem decidit que no haurien de formar part del conjunt de fenòmens que hauria de tractar el programari. Per evident que sembli aquesta darrera afirmació, convé tenir-la present, perquè sovint esperem que un programari tingui unes funcionalitats per a les quals no havia estat dissenyat.

### **3.2. Disseny i desenvolupament de l'arquitectura**

A l'hora de desenvolupar qualsevol sistema complex és imprescindible dividir-lo en diferents mòduls que puguin tractar-se com a elements independents responsables de tasques especialitzades i que alhora realitzin una tasca conjuntament que només s'aconsegueix fent-los treballar de manera interrelacionada.

L'arquitectura bàsica del verificador automàtic *El Corrector* es basa en l'existència de dos tipus de mòduls: els mòduls de processament i els mòduls de detecció. Els primers analitzen el text per descompondre'l i anotar-lo lingüísticament mitjançant tècniques estàndard de **processament de llenguatge natural** (PLN). Els segons revisen la seqüència d'elements lingüístics analitzats per detectar unitats o seqüències incorrectes que requereixin la generació de missatges d'error o d'avertiment.

Figura 1. Arquitectura del motor de correcció i generació de propostes de correcció emprada per *El Corrector*



A la figura 1 es mostra l'arquitectura utilitzada pel motor de correcció d'*El Corrector*. A l'esquerra es troben els mòduls de processament lingüístic, concretament:

- *Breaker*: aquest mòdul segmenta la seqüència de caràcters en diferents unitats: els paràgrafs, les oracions i els **tokens** (o elements textuais bàsics de processament lingüístic).
- *Labeler*: aquest mòdul consulta el diccionari i assigna a cada paraula la categoria morfosintàctica associada a la seva forma (o més d'una categoria si hi ha ambigüïtat); també inclou informació del registre, si escau, i de la variant a què pertany. Quan una paraula no és al diccionari la marca com a desconeguda.
- *Chooser*: aquest mòdul determina, en cas d'ambigüïtat, quina és la lectura (categoria) morfosintàctica més probable per a una paraula donades les paraules amb què concorre. Per exemple, donada la forma *casa* hauria de decidir si es tracta d'un nom femení singular o d'un verb en tercera persona singular del mode indicatiu i temps present.

#### Tokens

Es tracta essencialment de paraules, però també es tenen en compte signes de puntuació, espais, xifres, etc.

- *Other*: opcionalment l'arquitectura permet afegir altres mòduls que analitzin estructures de nivells lingüístics superiors, com locucions, estructures sintagmàtiques o entitats.

A mesura que es realitza el processament lingüístic els mòduls de la part dreta de la figura 1 van verificant la correcció de la seqüència de text enriquit amb informació lingüística. Concretament els tres mòduls són:

- *Typo*: aquest mòdul utilitza tècniques contextuais basades en regles sobre els aspectes gràfics (formals) dels *tokens* per determinar la utilització correcta dels espais, les majúscules i els signes de puntuació, i detecta errors ortotipogràfics.
- *Spell*: aquest mòdul utilitza tècniques no contextuais per determinar si les paraules apareixen al diccionari, i si formen part de la variant i registre corresponent, majoritàriament detecta errors ortogràfics.
- *Gram*: aquest mòdul utilitza tècniques contextuais basades en regles amb informació morfosintàctica per detectar seqüències corresponents a errors típics, i principalment detecta errors gramaticals.

Cada un dels mòduls verificadors genera una llista amb els errors trobats. Cada un dels errors inclou informació sobre la posició de l'element incorrecte, el tipus i codi d'error (que va associat sovint a un missatge d'error o advertiment), i les possibles correccions, sempre que hagi estat capaç de generar-ne. Aquestes llistes parcials proporcionades per cada un dels mòduls verificador s'integren en una llista única ordenada que és transferida a la interfície gràfica.

El resultat és filtrat segons les preferències de l'usuari o usuària final (variant, registre i tipus de correcció) i visualitzat segons un codi de colors per facilitar-ne la comprensió. A partir d'aquest punt, l'usuari o usuària pot revisar cada un dels errors i advertiments, confirmar-ne la validesa i, en cas necessari, triar entre les diferents correccions proposades.

### 3.3. Avaluació del sistema de verificació

Una vegada desenvolupat el sistema cal avaluar-ne el funcionament. En aquesta mena d'aplicacions el control de qualitat és fonamental, ja que una quantitat significativa de falsos positius o de falsos negatius faria l'eina totalment inservible. En el cas d'*El Corrector*, aquesta avaluació es va realitzar mitjançant el control de dos aspectes diferenciats: en primer lloc, l'acompliment de les especificacions funcionals i, en segon lloc, l'avaluació sobre una mostra de textos reals.

#### 1) Validació automàtica de les especificacions funcionals

Durant el desenvolupament de l'aplicació és important disposar d'un sistema d'avaluació automàtica que permeti garantir la bona direcció del projecte. Durant aquest període els diferents mòduls evolucionen i es milloren, els recursos lingüístics (diccionaris i regles) es modifiquen i s'amplien, i hi ha un risc real que els canvis interfereixin entre ells o amb altres mòduls i el funcionament general canviï. Pot millorar, que és allò que un esperaria, però també empitjorar, i llavors requeriria reconsiderar els canvis fets o fer-ne de nous; en qualsevol cas, cal un mecanisme de control.

Per això, tot just definides les especificacions lingüístiques és aconsellable crear un corpus d'errors, una base de dades amb centenars d'exemples de textos incorrectes amb l'error perfectament localitzat, classificat i amb la seva correcció indicada. Aquesta base de dades permet validar en tot moment la qualitat del verificador, i obtenir mesures precises i objectives del seu funcionament.

La idea és que un procés automàtic executi el verificador per a cada un dels casos (exemples) presents al corpus i analitzi si hi ha o no un error, el localitzi, el classifiqui i en generi la llista de propostes de correcció. Cada una d'aquestes informacions es compara amb les existents a la base de dades (al corpus) i, per tant, es pot saber exactament si l'error ha estat identificat i si la correcció proposada és correcta.

En finalitzar la validació automàtica s'obtenen un conjunt d'estadístiques globals, com els percentatges de precisió i cobertura del verificador o la quantitat de falsos positius i de falsos negatius. A més, es poden generar informes detallats sobre la precisió i la cobertura per a cada tipus i codi d'error.

En darrer terme, aquesta validació automàtica permet garantir (o si més no avaluar) la cobertura dels errors definits i, per tant, l'acompliment de les especificacions funcionals.

## 2) Validació manual sobre textos reals

Més enllà de l'acompliment de les especificacions, també és important i interessant conèixer la qualitat "real" del programari quan sigui utilitzat per usuaris o usuàries finals, és a dir, quan s'apliqui a textos diversos de diferent registre i amb una distribució no uniforme dels errors. En el cas d'*El Corrector* se'n va avaluar el comportament davant de diferents textos corresponents a publicacions científiques, notícies de premsa, textos de blocs i correus electrònics (Badia i altres, 2008).

A partir dels errors detectats pel verificador es van mesurar la precisió i la cobertura, així com l'adequació de la proposta de correcció. La cobertura es va mesurar com la proporció d'errors detectats entre els errors reals continguts en

els textos, i va resultar en un valor del 83%. La precisió es va mesurar com la proporció d'errors que efectivament ho eren entre tots els errors assenyalats pel programari, amb un resultat del 70%.

A més, aquesta avaluació va permetre fer una anàlisi dels punts flacs d'*El Corrector*: això va permetre saber que la majoria de falsos positius (57%) van produir-se per l'absència de paraules al diccionari, un fet previsible atès que es tracta d'un verificador normatiu i, per tant, no admet moltes paraules que sí s'admeten en àmbits específics. També que la majoria de falsos negatius (52%) es tractava d'errors que haurien necessitat una anàlisi lingüística més avançada (sintàctica o semàntica) per poder ser detectats i tractats correctament.



## Activitats

1. Prepareu un document amb força errors ortogràfics i apliqueu-hi el verificador ortogràfic del vostre processador de textos, i tracteu d'esbrinar les tècniques de correcció que fa servir per a elaborar la llista de suggeriments. Si voleu fer proves amb algun verificador que inclogui tècniques de detecció d'errors de base estadística podeu provar-ho amb l'MS Word 2007 fent servir l'anglès o el castellà (errors subratllats en vermell, verd i blau: per a més informació, vegeu [http://blogs.msdn.com/microsoft\\_office\\_word/archive/2006/10/25/contextual-spell-checking.aspx](http://blogs.msdn.com/microsoft_office_word/archive/2006/10/25/contextual-spell-checking.aspx)).
2. Què opineu que hauria de fer el verificador gramatical del vostre processador de textos amb les seqüències agramaticals però freqüents en català col·loquial com ara *Jo em sembla que no tens raó?* I amb les seqüències amb "desviacions semàntiques" com ara *La finestra llegia un conte?*
3. Comproveu si el verificador estilístic del vostre processador de textos inclou el càlcul d'algun índex de llegibilitat i proveu d'aplicar-lo a un document. Tracteu d'esbrinar-ne la fórmula utilitzada.

## Exercicis d'autoavaluació

1. Determineu a quina categoria d'errors ortogràfics pertanyen els exemples següents i raoneu la vostra resposta.
  - a) \**Creu el tap* per *treu el cap*.
  - b) \**Difícilment* per *difcilment*.
  - c) \**Apit* per *api*.
  - d) \**Satisfactori* per *satisfactori*.
2. Seguint les pautes de definició de regles per a la detecció d'errors mitjançant patrons, proposeu els patrons ampliatos d'errors per identificar i corregir errors gramaticals com els següents.
  - a) \**Vinc de el camp* per *vinc del camp*.
  - b) \**Hi han persones* per *hi ha persones*.
  - c) \**Hi ha que venir d'hora* per *cal venir d'hora*.
  - d) \**El fet de que vingui* per *el fet que vingui*.
3. Per què són difícils de preveure els errors gramaticals comesos en una llengua per les persones que la tenen com a llengua estrangera? Com afecta aquesta dificultat la verificació gramatical automàtica?
4. És raonable aplicar la fórmula de Flesch a textos escrits en llengua catalana? Justifiqueu la vostra resposta.

## Solucionari

### Exercicis d'autoavaluació

1.a) Podria tractar-se d'un error d'actuació relacionat amb els errors de la parla, per l'intercanvi del tret fonològic de localització entre les oclusives /t/ i /k/.

b) Error de competència per interferència amb les regles ortogràfiques del castellà. Formalment, és un error per substitució.

c) Error de competència per discrepància entre la parla i la normativa. Quant a la forma, es tracta d'un error per inserció.

d) Podria ser un error d'actuació resultat d'una distracció visual, per l'omissió de la segona lletra *s* en proximitat de la primera.

2.a) *de el > del*.

b) *hi han... NPL > hi ha... NPL*.

Aquest patró (en què NPL representa un nom en plural i els punts suspensius qualsevol seqüència de paraules) preveu la verificació d'errors com *\*hi han tres senyors*, amb paraules entre el verb presentador i el nom en plural.

c) *hi ha que VINF > cal VINF*.

d) *el fet de que > el fet que*.

3. Els errors gramaticals comesos en una llengua per les persones que la tenen com a llengua estrangera són difícils de preveure per la impossibilitat de tenir en compte totes les llengües del món i les seves possibles interferències. Per això, els verificadors gramaticals estan dissenyats per a la verificació dels errors gramaticals comesos per persones nadiues de la llengua, encara que també s'han elaborat verificadors gramaticals específicament adaptats per a la verificació dels errors comesos en una llengua per persones natives d'una altra llengua (per exemple, per verificar l'anglès escrit per estudiants alemanys de l'anglès com a segona llengua).

4. La fórmula de Flesh aplicada al català produeix uns resultats inadequats, perquè els paràmetres emprats per la fórmula (la longitud mitjana en síl·labes de les paraules i la mitjana de paraules per oració) s'han calibrat d'acord amb les pautes de la llengua anglesa.

## Glossari

**array associatiu** *m* sin. **hash**

**autòmat finit** *m* Un autòmat finit (o màquina d'estats finits) és un model matemàtic d'un sistema compost per estats, transicions i accions.

**cerca binària** *f* Algorisme informàtic que permet accelerar la cerca d'un element en una llista ordenada, en aquest cas, alfabèticament. Es basa a comparar l'element buscat amb el situat a la meitat de la llista; a partir del resultat es descarta una de les meitats de la llista. El procés es repeteix fins a localitzar l'element trobat.

**cobertura** *f* En sistemes de verificació, mesura de la quantitat d'errors que el sistema és capaç de detectar respecte al total que hi ha. En altres paraules, quants errors ha trobat del total que hi havia.

**corrector gramatical** *m* Eina informàtica, normalment integrada en un editor de textos, que s'utilitza per detectar errades gramaticals en un text escrit.

**corrector ortogràfic** *m* Eina informàtica, normalment integrada en un editor de textos, que s'utilitza per detectar errades ortogràfiques en un text escrit.

**etiquetador** *m* En l'àmbit de la lingüística de corpus, és un programa informàtic que permet l'assignació automàtica d'una *etiqueta* (*tag*, en anglès) morfosintàctica amb de la seva categoria gramatical a cada paraula d'un text. *en tagger*

**fals positiu** *m* En sistemes de correcció, un element correcte marcat pel corrector com a incorrecte, és a dir, un *fals error*.

**fals negatiu** *m* En sistemes de correcció, un element incorrecte no marcat pel corrector, és a dir, una detecció fallida.

**forma** *f* Cada una de les formes flexionades que poden obtenir-se a partir d'un paradigma flexiu. Per exemple: *nen, nena, nens i nenes*, són les formes corresponents a la flexió del lema *nen*.

**formari** *m* Diccionari de formes, llista o base de dades que recull el conjunt de paraules flexionades d'una llengua. Per exemple, les formes masculines i femenines, les formes singulars i plurals, o tots els elements presents en les conjugacions verbals.

**hash** *m* En informàtica, estructura de dades que permet indexar de manera molt eficient una informació. Es basa a indexar una dada a partir d'un valor 13èric obtingut a partir de la mateixa informació, garantint que no hi puguin haver dues dades diferents que comparteixin el mateix índex.

sin **array associatiu** o **llista associativa**

**lema** *m* Forma arbitrària seleccionada per referir-se a una conjunt de flexions. Habitualment es pren com a lema la forma singular masculina dels noms i dels adjectius o l'infinitiu dels verbs. Per exemple: *nen*, és el lema de les formes *nen, nena, nens i nenes*; i *jugat* és el lema de les formes *jugo, jugues, jugava, jugant, jugaríem...*

**lemari** *m* Diccionari de lemes, llista o base de dades que recull el conjunt de lemes d'una llengua. Per exemple, les formes masculines singulars dels noms i adjectius, els infinitius dels verbs...

**lexicó** *m* Llista de paraules ordenades amb informació lingüística associada (lema, categoria gramatical, gènere, nombre, etc.).

**llista associativa** *f* sin. **hash**

**mètrica de llegibilitat** *f* En el camp de l'estilometria, equació que permet estimar el grau de dificultat de lectura d'un text en funció de les seves característiques superficials.

**model estilístic** *m* En sistemes de verificació, idealització d'una variant lingüística geogràfica, social o contextual, utilitzada per agrupar un conjunt de regles, elements lèxics i valors de llegibilitat, per prendre'l com a referència a l'hora de verificar l'estil d'un document.

**precisió** *f* En sistemes de verificació, mesura de la quantitat d'errors autèntics de tots els que el sistema ha indicat. En altres paraules, quants dels errors trobats eren realment errors.

**reconeixement de patrons** *m* Tècnica informàtica utilitzada en el camp del processament del llenguatge natural, que serveix per detectar les seqüències de paraules d'un text que responen a unes pautes configuracionals determinades.

**registre** *m* Varietat funcional d'una llengua definida d'acord amb els factors de la situació comunicativa, com ara el nivell de formalitat, el tema del qual es parla, etc.

**tagger** *m* Vegeu **etiquetador**.

**text no restringit** *m* Expressió per referir-se a textos de qualsevol tipus (origen, format, registre, variant dialectal, etc.). Es fa servir en oposició a l'expressió *textos de domini o controlats*, que vol dir que són textos amb un ús uniforme de la llengua.

**token** *m* En processament de llenguatge natural, unitat mínima d'anàlisi. Tot i que s'associa a una paraula, també es consideren *tokens* els signes de puntuació i els elements compostos com nombres decimals, dates, telèfons...

**TRIE** *m* En informàtica, estructura de dades que permet indexar de manera molt eficient dades seqüencials. S'utilitza habitualment per emmagatzemar paraules mitjançant la creació d'un arbre jeràrquic, de manera que les paraules que comparteixen prefixos, s'emmagatzemen en branques properes.

**variant** *f* Expressió lingüística diferent d'una altra per la forma.

**varietat** *f* Ús específic que es fa d'una llengua d'acord amb la procedència geogràfica, històrica o social dels parlants o amb la funció comunicativa, i que es caracteritza per una determinada concurrència de variants lingüístiques.

**verificació lingüística** *f* En el camp de l'escriptura assistida per ordinador, terme general que inclou la verificació ortogràfica, la verificació gramatical i la verificació estilística d'un text. Els programes informàtics que la duen a terme s'anomenen, respectivament, *verificadors ortogràfics*, *verificadors gramaticals* i *verificadors estilístics*. El terme *verificador* és més genèric que el de *corrector* i no implica necessàriament la correcció dels problemes lingüístics detectats.

## Bibliografia

**Costa Carreras, J.; Nogué Serrano, N.** (coord.) (2006). *Curs de correcció de textos orals i escrits: pràctiques autocorrectives* (3a. ed. rev.). Vic: Eumo Editorial.

**Damerau, F. J.** (1964, març). "A technique for computer detection and correction of spelling errors". A: *Communications of the ACM* (vol. 7, núm. 3, pàg. 171-176).

**Gamon, M.; Leacock, C.; Brockett, C.; Dolan, W. B.; Gao, J.; Belenko, D.; Klementiev, A.** (2009, juny). "Using Statistical Techniques and Web Search to Correct ESL Errors". *CALICO Journal* (vol. 26, núm. 3).

**Gómez Guinovart, J.** (1999). *La escritura asistida por ordenador: problemas de sintaxis y de estilo*. Vigo: Universidade de Vigo (Servicio de Publicacions).

**Granger, S.** (2003). "Error-tagged Learner Corpora and CALL: A Promising Synergy". *CALICO Journal* (vol. 20, pàg. 465-480).

**Leacock, C.; Chodorow, M.** (2003). "Automated grammatical error detection". A: M. D. Shermis, J. Burstein (editors). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

**Mitton, R.** (1996). *English Spelling and the Computer*. Londres: Longman.

**Quixal, M.; Badia, T.; Benavent, F.; Boullosa, J. R.; Domingo, J.; Grau, B.; Massó, G.; Valentín, O.** (2008, maig). "User-Centred Design of Error Correction Tools". A: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marràqueix, el Marroc.

