

## Presentació

Quan a SoftCatalà vàrem decidir fer la localització de l'OpenOffice, em vaig oferir a cuidar-me'n del corrector ortogràfic. Com sabeu tots els qui heu col·laborat alguna vegada en coses d'aquestes, oferir-te a fer una feina i que te l'assignin és automàtic. Així que ja em teniu, com els altres companys en les seves tasques respectives, enterrat davall una allau d'informació, intentant fer funcionar alguna cosa que s'assembla a un corrector en català.

El corrector d'OpenOffice, batejat com MySpell, s'assembla bastant a ispell, un corrector ben conegut i molt usat sobretot en Linux. Per això vaig agafar els fitxers de català per a ispell, fets per Ignasi Labastida, els vaig adaptar al format requerit per MySpell, vaig trepitjar els fitxers de l'anglès i vaig fer la primera prova. Resultat: el programa penjat. Un sospitós habitual: potser la llista de paraules catalana era massa llarga comparada amb l'anglesa i el programador no havia reservat prou espai. No era això; els problemes es repetien amb una llista molt més curta. Un altre sospitós de sempre, els caràcters amb codi alt: se suprimeixen els accents, les ces trencades, els punts volats i la cosa comença a funcionar.

La primera passa és molt modesta, però de qualche manera s'ha de començar. Les proves varen servir per constatar que tant el programa de correcció com el corpus de paraules reconegudes s'havien de millorar.

## Procés de correcció

Un corrector ortogràfic és un programa relativament autònom que atén consultes d'un altre programa, per exemple un editor de textos.

Mentre l'usuari va escrivint (o una vegada fet el document), l'editor va trossejant el text en paraules i les sotmet al corrector. Per a cada paraula el corrector diu si és correcta o no. Com que es passa la paraula sense gens d'informació sobre el seu entorn, només es poden detectar paraules que per si mateixes estiguin mal escrites, és a dir, no es poden corregir errades que afectin més d'una paraula. Si la paraula resulta ser incorrecta, l'editor pot demanar al corrector una llista de substituïts, segurament per deixar-ne triar una a l'usuari.

MySpell pot corregir en diverses llengües. Hi ha una part comuna a totes, que es pot anomenar “motor de correcció”, i diversos fitxers amb les dades específiques de cada llengua. En el cas del català es varen fer modificacions al motor, per millorar-ne les prestacions i per solucionar alguns problemes deguts a l'excessiva orientació cap a l'anglès del motor original.

El corrector disposa d'una llista amb totes les paraules correctes. En el cas del català i de totes les llengües amb molta flexió nominal i verbal, no es pot mantenir una llista exhaustiva, sinó que es fa servir una llista d'arrels i formes irregulars, i una altra llista, molt més breu, de regles per generar les diverses formes. Les dues funcions del corrector consisteixen, aleshores, a decidir si una determinada paraula és a la llista de totes les formes, per una banda, i a obtenir la llista de les formes que estan a una “distància” mínima de la paraula. Mentre que decidir si una paraula forma part de la llista és fàcil d'implementar, cercar substituïts per a una paraula incorrecta és molt més difícil, començant pel mateix concepte de “distància” i acabant per una implementació eficient.

La qualitat d'un corrector s'ha de mesurar segurament fent servir dos paràmetres: quantes paraules correctes reconeix i en quin percentatge de casos proposa la paraula correcta. És evident que aquest segon criteri és més difícil d'avaluar però sí que es pot fer una estimació relativa comparant la resposta de diversos correctors per a les mateixes paraules. S'ha de dir que MySpell, en aquest aspecte, queda enrere respecte de la “competència” per a moltes paraules, però també que l'altre falla en paraules que el nostre resol perfectament. Per exemple, propostes per a “colecció”, amb ella simple: el nostre, “col·lecció”; l'altre, “col·lació”. La qualitat de les

propostes i els algorismes per implementar-ho eficientment és un tema obert de debat a OpenOffice.

El corrector MySpell, com fa ispell, cerca propostes de substitució modificant la paraula original, comprovant si del canvi en resulta una paraula correcta i tornant les noves paraules que resulten ser a la llista. Per generar aquestes paraules de prova, aplica les següents manipulacions: afegir una lletra, suprimir una lletra, canviar d'ordre dues lletres consecutives i dividir la paraula en dos trossos (per si s'hagués deixat l'espai). Aquest algorisme no fa propostes evidents per a algunes errades comunes en català, com ara posar *ela* geminada per comptes d'*ella* simple, o al revés. Ja veurem de quina manera ho vàrem millorar en la versió catalana. Aquesta millora va ser implementada independentment pel responsable de MySpell, a proposta nostra i d'usuaris d'altres llengües, i apareixerà a la versió 1.1 d'OpenOffice. Tot i el guany de qualitat que s'ha aconseguit en les propostes, és evident que caldria trobar un algorisme molt més eficaç. Sovint, especialment amb paraules llargues escrites incorrectament, una de les paraules correctes s'assembla molt més a la dolenta més que qualsevol altra. Idealment, en aquest cas aquesta paraula s'hauria de proposar, cosa que moltes vegades no passa.

## **Adaptació de BreakIterator**

En el primer pas del procés de correcció (la divisió en paraules) ja hi havia algunes dificultats. El punt volat no era reconegut com un caràcter interior d'un mot, sinó com a separador. Alhora que es va modificar la qüestió del punt volat es va decidir que els guionets també es considerarien caràcters catalans, ja que formen part de certes paraules compostes, i també perquè així es podrien corregir les combinacions de verbs i pronoms, un punt que té fama (crec que immerescuda) d'enrevessat. També els guions llargs tenien un tractament inadequat per al català, ja que es consideraven part de la paraula, en contra del costum establert de fer-los servir de separadors. Igualment es va modificar aquest aspecte.

El resultat de les modificacions va ser un dels múltiples components d'OpenOffice, que forma part actualment de la nostra distribució d'aquest programa, reemplaçant el component original.

## **Adaptació de MySpell**

Una vegada fetes les modificacions al divisor de paraules, l'editor enviava les paraules senceres a corregir (incloent els apòstrofs i els guionets interiors i els punts volats). Tanmateix, a l'espera de tenir una llista de paraules i de regles adequada, es detectaren alguns errors i detalls que calia millorar. Tots els canvis es varen notificar al responsable del projecte MySpell. Alguns varen ser acceptats i han passat o passaran a formar part de la versió “general” de MySpell. Altres no varen prosperar, i es va optar per fer una versió pròpia de MySpell, la que forma part de la nostra distribució. Finalment, altres no es varen acceptar, però han reaparegut després, proposats des d'altres punts del món.

Alguns dels errors afectaven les paraules escrites totalment o parcialment en majúscules. Es marcaven com a faltes moltes d'aquestes paraules amb guionets o apòstrofs i els noms propis amb un article o una preposició prefixats en majúscula. També, en casos d'errors genuïns no s'arribava a proposar la paraula correcta. La solució d'aquests errors va obligar a la modificació del codi font del corrector.

Un canvi encara més important afectà notablement la qualitat dels suggeriments. En la versió original, per a la paraula “aixo”, escrita sense accent, no es proposava la paraula amb accent, evidentment la més acostada. El problema no era que el corrector no la generàs, sinó que estava molt avall dins la llista, i ni tan sols s'arribava a mostrar. Per a un altra errada

habitual, posar la geminada per la simple, o al revés, no es proposava el substitut natural, ja que la paraula correcta i la incorrecta estaven separades per dues lletres. Per tot això, es va pensar de crear una taula de substitucions preferents, per a caràcters o grups de caràcters, que s'aplicava abans dels altres mètodes per cercar la paraula correcta. Com a conseqüència, el nostre corrector ara proposa “això” com a primera alternativa per a la mateixa paraula sense accent, i “col·laborar” per a la paraula sense la geminada. Aquesta modificació del codi va ser més profunda que l'anterior. La futura versió de MySpell també ho farà servir, amb una implementació diferent, però amb una total compatibilitat dels nostres fitxers. El sistema és molt potent, i possiblement encara no se n'ha tret tot el profit.

## **Llistes de paraules i regles**

Amb el motor de correcció afinat, faltaven encara les llistes d'arrels i de regles per al català. Per raons d'eficàcia, per comptes de tenir una llista amb totes les paraules possibles, se'n fa servir una amb les arrels i les formes irregulars, i una altra amb regles de derivació morfològica i informació extra per al motor. Aquesta organització en dos fitxers és heretada d'ispell.

El primer candidat per als fitxers eren els que hi havien d'ispell per al català, una mica modificats. Aviat es va comprovar que tenien diversos inconvenients. No eren gaire complets, ni pel que fa a les paraules que reconeixien ni pel que fa a les regles de derivació, com reconeixia un comentari dins el mateix fitxer, a més de tenir errors, alguns de bastant severos. Hi mancaven gairebé per complet els paradigmes verbals valencians i balears. També hi faltava informació sobre la majoria dels mots que poden dur un article o un pronom apostrofat. Es va considerar que seria més fàcil fer llistes noves que adaptar les existents.

Es va començar pel fitxer de regles morfològiques. Una primera decisió va ser incorporar-hi, en pla d'igualtat, els paradigmes dels grans dialectes, amb molta variació sobretot pel que fa als verbs. No es va crear directament el fitxer de regles sinó un altre amb més informació, a partir del qual es va generar aquell. Una idea inicial era fer tres versions diferents del corrector, una per cada gran dialecte, amb les mateixes paraules però amb les regles morfològiques pròpies de cada zona. Al capdavant es va descartar aquesta idea i ara hi ha llistes comunes, amb un cert detriment de la qualitat del corrector. El fitxer de regles conté unes nou-centes entrades, com pertoca a una llengua romànica. S'ha de tenir en compte que juntament amb les terminacions verbals hi ha les nominals, a més dels articles, els pronoms personals prefixats i totes les combinacions de pronoms sufixats.

La llista d'arrels es fa normalment amb un programa auxiliar que “comprimeix” una llista exhaustiva de paraules. En el nostre cas, vàrem optar per generar directament la llista amb un programa escrit expressament en Perl. Es varen agafar totes les entrades del Diccionari de l'Institut d'Estudis Catalans, diverses llistes de noms propis (topònims i antropònims) i altres llistes de noms comuns (sobretot gentilicis). A partir d'aquestes dades es va crear el fitxer. El programa està completament orientat al català. A tall d'exemple, conté codi per decidir si una paraula pot dur o no article, en funció de la categoria, de si comença o no per vocal, del gènere i, eventualment, de si la primera síl·laba és tònica o àtona. La llista generada consta d'un 170.000 entrades.

## **Productes derivats**

Com a resultat de la feina feta amb el corrector, i aprofitant el gran parentiu entre MySpell i ispell, es varen crear fitxers per a aquest darrer programa. Contenen les mateixes paraules, amb l'única excepció de les abreviatures, per culpa del tractament que fa ispell dels punts, completament diferent del que fa MySpell. Els fitxers d'ispell formen part de la distribució

d'AbiWord, i es poden fer servir amb ispell o amb qualsevol altre programa que faci servir les mateixes dades. Creim que actualment són els fitxers per a ispell més complets que es poden trobar en català.

## **Per fer**

El corrector està fet i és perfectament funcional. He escrit aquesta ponència fent-lo servir i només m'ha marcat com a faltes els noms de programes (inclòs el seu mateix nom, MySpell!) i algunes paraules que he posat com exemple d'incorreccions. Tanmateix, encara hi ha molt d'espai per treballar. Aquí teniu algunes propostes:

S'ha d'ampliar la llista de paraules. Hi falten molts de noms propis (per exemple, marques comercials, cognoms, noms de persona) i molts de comuns (com ara neologismes, gentilicis, etc.).

S'ha de millorar la qualitat dels suggeriments. Com hem dit, és una qüestió que s'està debatent actualment a OpenOffice. És necessari estudiar què fan els altres correctors (siguin de català o d'altres llengües) per millorar el nostre. Per exemple, alguns correctors, com aspell (pensat per succeir ispell), fan servir algorismes basats en metàfons, una espècie de transcripció fonològica de les paraules. S'hauria de tenir un corpus amb errades típiques per tal de poder fer comparacions entre diversos correctors, o entre diverses versions.

No tots els usuaris necessiten el mateix corrector. Aquest hauria de ser parametrizable, segons les preferències, per exemple geogràfiques, de cadascú. També s'haurien de poder activar o desactivar vocabularis especialitzats, que evidentment estan per fer.

Hi ha molts de programes mancats de corrector ortogràfic, o amb correctors de poca qualitat. El corrector podria ser un programa independent, que oferís servei a tota una sèrie de clients. Per a cadascun d'aquests només s'hauria d'escriure un mòdul de connexió, generalment més fàcil i més lleuger que integrar-hi tot el corrector. A més, es garantiria que tots els programes fessin servir la versió més actual del corrector i les millores (per exemple, les paraules que cada persona volgués afegir al seu diccionari particular, o les noves versions), estarien instantàniament a disposició de tots els clients.

## **Altres plans**

Vaig dedicar molt de temps i molts d'esforços al corrector ortogràfic, però això ja és vell. Darrerament, estic treballant sobre un corrector gramatical, per dir-ho pomposament, o un corrector de frases, per dir-ho més planer. La feina és molta, però els primers resultats són esperançadors. El principi del corrector seria el següent: donada una frase, se'n cerquen les possibles anàlisis, fent durant el procés canvis eventuais, per exemple de concordança; en tenir una anàlisi convenient, es regenera la frase tenint en compte les anotacions sobre errades i els canvis. Aquest tipus de correctors no tenen gaire bona anomenada, però crec que són particularment útils per a llengües com la nostra, en què abunden les faltes.

Considerau per exemple aquestes tres frases:

“El nins juguen al carrer”. L'anàlisi amb l'article en singular no arriba enlloc. Una regla substitueix el singular pel plural, la qual cosa condueix a una anàlisi satisfactòria. Es reconstrueix la frase “Els nins juguen al carrer”, que es proposa com a correcció amb l'explicació corresponent.

“L'insigne ciutat d'Alcúdia”. He agafat aquest exemple, justament perquè no és detectat per l'únic corrector gramatical popular de català. El problema és que “insigne” pot ser tant masculí com femení, però aquí és femení, ja que concorda amb “ciutat” i, per tant, se li aplica la regla que mana no apostrofar. En aquest cas, l'anàlisi es faria sense problemes, però durant la

regeneració de la frase es detectaria l'errada. La frase demostra com no és suficient una anàlisi que només tingui en compte grups de dues paraules consecutives.

“El tamany de la imatge és excessiu”. En aquest cas, l'anàlisi aniria bé, gràcies a la presència del barbarisme “tamany” a la llista de paraules, evidentment amb una anotació en el sentit que ha de ser substituït, per exemple, per “mida”, que no té el mateix gènere que “tamany”. Durant la regeneració, es forçaria l'aplicació de les regles de concordança de les frases copulatives, i el resultat seria, per tant, “La mida de la imatge és excessiva”.

Una altra idea que em ronda pel cap, relacionada amb tot el que he exposat, és la creació d'una espècie de “portal” d'eines i dades lingüístiques per al català. Crec que molts dels recursos sobre el català estan massa dispersos i s'haurien d'agrupar. Exemples del material que hi podria haver són: algorismes d'ordenació lexicogràfica, de divisió sil·làbica, de conversió de números a lletres o de transcripció fonètica, corpus de noms propis i comuns, de regles gramaticals o d'errades freqüents, documentació, des de morfologia fins a APIS de programes.

## **Col·laborar**

Si voleu, podeu col·laborar. Hi ha feina a voler per a tothom. El més senzill és trobar errades al que ja hi ha fet o proposar millores. Una mica més complicat, però encara a l'abast de tothom, és fer llistes, tant de paraules bones com d'errades, de frases o regles que es puguin fer servir per construir i provar l'eventual corrector gramatical, i sobretot estudiar amb esperit crític altres productes que ens caiguin a les mans. Més difícil, però més gratificant, és fer o adaptar programes. Es pot intentar aplicar el que hi ha fet a programes com editors de texts o de pàgines web, compaginadors, clients de correu o de xat i tot el que vulgueu. Finalment, es pot explorar terreny desconegut. El corrector gramatical en podria ser un bon exemple.