DIGITHUM
Les humanitats en l'era digital
www.uoc.edu/digithum

Un corrector gramatical basat en cerques per Internet

Joaquim Moré
Investigador de l'Internet Interdisciplinary Institute (IN3) de la UOC
jmore@uoc.edu

Data de presentació: gener del 2006
Data de publicació: maig del 2006

CITACIÓ RECOMANADA
MORÉ, Joaquim (2006). «Un corrector gramatical basat en cerques per Internet». *Digithum* [article en línia]. Núm. 8. UOC. [Data de consulta: dd/mm/aa].
<http://www.uoc.edu/digithum/8/dt/cat/more.pdf>
ISSN 1575-2275

Resum
En aquest article presentem un corrector gramatical de l'anglès destinat a escriptors no angloparlants. La principal característica d'aquest corrector és l'ús d'un motor de cerca per Internet. Com que hi ha un gran nombre de pàgines web escrites en anglès, el sistema fa la hipòtesi que un segment de text que no és present en cap pàgina web és probablement un segment de text mal escrit. El sistema també fa la hipòtesi que a la Xarxa hi trobarà exemples que ensenyaran a l'usuari com ha d'expressar el contingut del segment de text d'una manera gramatical i idiomàtica. Per tant, un cop el corrector avisa l'usuari que és millor verificar un segment del seu text, el motor cerca contextos que poden ser útils a la persona que escriu a l'hora de decidir si corregeix el segment o no. Gràcies també a l'ús d'un motor de cerca, el corrector suggereix a l'escriptor que utilitzi expressions que són més freqüents a la Xarxa en comptes de l'expressió que ha escrit.

Paraules clau
correcció gramatical, correcció estilística, processament del llenguatge natural

Abstract
This paper presents an English grammar and style checker for non-native English speakers. The main characteristic of this checker is the use of an Internet search engine. As the number of web pages written in English is immense, the system hypothesises that a piece of text not found on the Web is probably badly written. The system also hypothesises that the Web will provide examples of how the content of the text segment can be expressed in a grammatically correct and idiomatic way. Thus, when the checker warns the user about the odd nature of a text segment, the Internet engine searches for contexts that can help the user decide whether he/she should correct the segment or not. By means of a search engine, the checker also suggests use of other expressions that appear on the Web more often than the expression he/she actually wrote.

Keywords
grammar checking, style checking, natural language processing

1. Introducció

El corrector gramatical que presentem en aquest article es desenvolupa a la Universitat Oberta de Catalunya. El seu objectiu principal és ajudar el personal docent de la institució i els seus investigadors que no són angloparlants a escriure textos en anglès (articles, missatges de correu electrònic, etc.). Malgrat tenir un domini acceptable de la llengua, la majoria no se senten prou segurs de la correcció dels textos que escriuen i els sembla que moltes de les frases dels seus escrits delaten un nivell d'expertesa insuficient, perquè no són prou idiomàtiques. Ara bé, se senten segurs del que escriuen quan veuen les seves frases o els seus segments de text en un document ja publicat en anglès, sempre que la correcció gramatical i estilística del document estigui garantida. Si no troben la frase o segment en cap document, la inferència que hi hagi un error només es justifica si el nombre de documents disponibles és elevat i els documents són variats. A Internet hi ha un nombre immens de documents, de tipus i gèneres molt variats; per la qual cosa, la principal característica d'aquest corrector és que usa un motor de cerca per Internet per a detectar segments de text de l'escrit de l'usuari que no es troben en cap pàgina web. Per a cada un d'aquests segments, el corrector informa l'usuari que el segment és nou (*brand-new*) en l'univers d'Internet i que probablement això passa perquè està mal escrit. La probabilitat que això sigui efectivament així és prou alta, tenint en compte que l'escriptor no és angloparlant i que no té un coneixement molt profund de la llengua. Després, el corrector cerca pàgines web que contenen diferents maneres d'expressar el contingut del segment (variants) i mostra a l'usuari contextos amb aquestes variants a partir de la pàgina de resultats de la cerca.

L'evidència a partir de corpus grans s'ha aplicat en el camp de la generació del llenguatge natural per a escollir una entre vàries realitzacions possibles d'una frase (Langkilde i Knight, 1998; Langkilde, 2002) i també s'han utilitzat motors de cerca per Internet per a avaluar les regles de detecció d'errors d'alguns correctors gramaticals (Naber, 2003). El corrector basat en corpus que aquí presentem mai no diu a l'usuari com ha d'escriure; aniria en contra de l'ús creatiu del llenguatge si jutgés un segment com a incorrecte perquè no el troba a la Xarxa. El corrector simplement avisa l'escriptor i mostra contextos que contenen les variants del segment que ell ha escrit i que el motor de cerca per Internet ha trobat. Aquests contextos es consideren útils per a l'usuari, el qual s'adona dels seus errors gramaticals i estilístics, i l'ajuden a decidir de rescriure el text o, en canvi, deixar-lo tal com està si en els contextos no hi veu cap indici que ho aconselli.

2. Descripció dels components

El corrector gramatical consta dels components següents:

- Interfície d'usuari
- Etiquetador
- Analitzador de fragments (*chunker*)
- Motors de cerca d'Internet
- Detector de fragments nous
- Detector de fragments millorables
- Cercador i visualitzador d'exemples

Interfície d'usuari

La interfície d'usuari carrega el document que l'usuari vol verificar (ara per ara el document ha d'estar en format .txt). L'usuari pot comprovar un fragment de text concret seleccionant-lo i fent-hi clic. En aquest cas el sistema verifica el segment seleccionat. Si l'usuari no selecciona cap segment, el sistema verifica tot el text.

Etiquetador

L'etiquetador etiqueta cada paraula d'una cadena segons la seva categoria gramatical. L'etiquetador que el sistema utilitza és la versió demo del TreeTagger (Schmid, 1994) per a Windows.[www1] La versió demo no pot etiquetar més de dues-centes paraules. De totes maneres, com que s'ha treballat sobretot en la verificació de fragments seleccionats per l'usuari, en aquest cas és difícil que el nombre de paraules superi aquest límit. La sortida de l'etiquetador és una llista de paraules etiquetades que tenen el format següent: paraula-categoria gramatical-lema.

Analitzador de fragments

L'analitzador de fragments agrupa en segments les paraules d'un fragment de text etiquetades segons la seva categoria gramatical. En aquest moment hem establert els segments següents:

- *Nominal*: cadena de paraules que són determinants, adjectius o noms i que formen un sintagma nominal (per exemple, *an Internet search engine*).
- *Verbal*: cadena de paraules que formen un verb simple o compost.
- *Verbal+nominal*: cadena de paraules que conté un segment verbal seguit d'un de nominal (per exemple, *organise the academic activity*).
- *Nominal+preposició+nominal*: cadena de paraules que conté dos segments nominals lligats per una preposició (per exemple, *laborer on a farm*).

[www1]: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

- *Verbal+preposició+nominal*: cadena de paraules que conté un segment verbal i un de nominal lligats per un preposició (per exemple, *carry out a project*).
- *Preposició+nominal*: cadena de paraules que conté una preposició seguida d'un segment nominal. La cadena no és dins d'un segment més llarg (per exemple, *on the one hand*).
- *Adverbial+verb/adjectiu*: cadena de paraules que conté un segment adverbial i un de verbal o un d'adjectival (per exemple, *also display examples*).

Els segments contenen conceptes i relacions entre conceptes. Considerem les preposicions i els verbs com a paraules que relacionen conceptes.

Motors de cerca per Internet

El corrector utilitza el motor de Wordnet 2.0,[www2] una base de dades lèxico-semàntica disponible en línia, per a obtenir informació sobre com es poden expressar els conceptes. Els motors que s'utilitzen per a obtenir el nombre de pàgines web que contenen un segment de text (resultats de la cerca) són el motor de cerca de Yahoo[www3] i el d'Altavista.[www4]

Detectors de segments nous i millorables

A partir de la pàgina de resultats de la cerca, els detectors de segments nous i millorables saben si el segment és nou (no s'ha trobat cap coincidència en cap pàgina web). Si no ho és, els detectors també jutgen si el segment es pot millorar (segment millorable).

Cercador i visualitzador d'exemples

Quan un segment és nou o es considera que és millorable, el cercador busca a la Xarxa pàgines web que contenen variants d'aquest segment i mostra contextos (*snippets*) amb aquestes variants a partir de la pàgina de resultats de la cerca. Aquests contextos poden ser útils per a l'usuari a l'hora de decidir si reescriu el contingut del segment o no. El nombre màxim de contextos que es poden mostrar en una pàgina de resultats és de cent.

3. Detecció de segments nous i millorables

Els segments nous són els que el motor de cerca busca a la Xarxa i que com a resposta obté una pàgina de resultats amb la seqüència «We didn't find any web pages» ('No s'ha trobat cap pàgina web'), o bé que a la pàgina de resultats de les cent primeres pàgines web trobades la coincidència exacta no està marcada amb negreta. La detecció de segments improbables és més complexa.

3.1. Wordnet i la detecció de segments millorables

El detector de segments millorables activa el motor de cerca de Wordnet per a trobar millors maneres d'expressar el contingut d'un segment, que anomenem *segment hipòtesi*. Per exemple, quan el segment hipòtesi és del tipus *preposició+nominal*, el detector fa la hipòtesi que aquest segment de text és una manera d'expressar un concepte o és un connector discursiu. D'acord amb l'organització de Wordnet, el motor cerca els *synsets* del nucli nominal (cada *synset* és un conjunt de paraules sinònimes que denota un concepte) i també les glosses que expliquen cada un d'aquests sentits. S'etiqueta i se segmenta cada glossa on apareix el nucli i es compara el segment que el conté amb el segment hipòtesi. Si els segments coincideixen en tipus però varien per una paraula funcional, el segment hipòtesi es considera millorable. Vegem-ne un exemple. Imaginem-nos que l'escriptor ha escrit:

(1) *In the one hand, we explain the antecedents in the study of the cognitive processes…*

In the one hand no és nou, però el motor de cerca de Wordnet troba *on the one hand…, but on the other hand…* en la glossa del sentit 7 de 'hand'. Després d'haver etiquetat i segmentat la glossa, el detector s'adona que *on the one hand* forma el mateix segment sintàctic que *in the one hand*, que no apareix en la pàgina de resultats de Wordnet. Per tant, el corrector mostra el missatge següent:

(2) **hand** - (one of two sides of an issue; *on the one hand…, but on the other hand…*)

Aquest missatge és la glossa del sentit 7 de 'hand' a Wordnet i pot ser útil perquè l'usuari s'adoni que hauria de revisar *in the one hand*.

3.2. Aprofitar els «Did you mean…?»

Quan a la pàgina de resultats de la cerca hi apareix la pregunta «Did you mean…?» ('Volies dir…?'), la forma suposada també s'etiqueta i se segmenta per a comprovar si la seva estructura sintàctica coincideix amb la del segment hipòtesi. Si és així, se

[www2]: http://www.cogsci.princeton.edu/~wn
[www3]: http://www.yahoo.com
[www4]: http://www.altavista.com

cerca la forma suposada i es compara el seu nombre de resultats amb el nombre de resultats de la hipòtesi. Es considera que el segment hipòtesi és millorable quan el seu nombre de resultats és més petit. Per exemple, imaginem-nos que l'usuari escriu:

(3) *…it displays real-English examples with an Internet searcher.*

La pàgina de resultats per a *Internet searcher* conté la pregunta «Did you mean 'Internet search'?» ('Volies dir *Internet Search?*'). S'etiqueta *Internet search* i s'identifica com un segment nominal, igual que *Internet searcher*. El nombre de resultats d'*Internet searcher* (1.660) es compara amb el d'*Internet search* (3.220.000). Segons aquesta comparació, *Internet searcher* es considera millorable.

3.3. Detecció de la variant més freqüent

La variant d'un segment pot ser una cadena amb les mateixes paraules però en un ordre diferent de com apareixen en el segment hipòtesi. Vegem per exemple:

(4) *…in order to detect odd pieces of text and to also display helpful contexts.*

Si l'usuari vol verificar *and to also display*, l'adverbi *also* es col·loca en la posició més extrema a l'esquerra i es fan noves crides amb el motor de cerca movent l'adverbi una posició en cada crida de dreta a esquerra. El motor busca cada variant i el detector compara el nombre de resultats (*also and to display*: 0; *and also to display*: 340; *and to also display*: 13; *and to display also*: 2). Com que els resultats de *and to also display* només superen *also and to display* i *and to also display*, aquest segment es considera millorable.

4. Mostra de contextos útils

Quan un segment es considera millorable, el corrector mostra contextos curts extrets de les pàgines web que contenen la forma alternativa escollida. Aquests contextos són els que apareixen a la pàgina de resultats (*snippets*). La variant escollida apareix en negreta. Per tant, en el cas d'*Internet search*, el sistema mostra contextos com ara (5i) i (5ii).

(5) i) *…***Internet Search** *Tools. Single SearchEngines/Portals…*

ii) *With billions of pages on the Web, you use a search engine if you're looking for something specific. Learn how search engines acquire, store and organize all that data to help you find what you're […] like most people, you visit an* **Internet search** *engine.*

Després d'haver llegit (5ii), l'usuari que ha escrit *Internet searcher* probablement preferirà escriure *Internet search engine*. Aquest és un exemple de com el sistema pot ser útil per als traductors, qui han de manejar terminologia.

En el cas que hem vist de *and also to display* es mostraran contextos com ara (6):

(6) *…Sometimes the use of a spreadsheet can help the pupils to perform calculations more easily and* **also to display** *their results graphically in the form of bar charts and pie charts. This facility to…*

Pel que fa a segments nous, la cerca de contextos útils es fa substituint les paraules que relacionen conceptes per un nou element. Quan el segment és de tipus *verbal+nominal*, se substitueix el verb per un dels seus sinònims. El sinònim pertany als *synsets* del verb segons la pàgina de resultats del motor de Wordnet. Llavors, els motors de Yahoo i Altavista cerquen documents amb les noves paraules clau. Si es troben contextos, aquests es mostren a l'usuari. Per exemple, si l'usuari escriu el segment nou *to devise the academic activity*, 'devise' se substitueix per un sinònim de Wordnet diferent ('organise', 'organize', 'machinate'…) en *n* cerques, essent *n* el nombre d'elements dels *synsets* del verb. Llavors es mostren contextos com ara (7).

(7) *Committees including the important General/Professorial/Academic Board, and the Finance Committee […] and lectureships, and* **organise the academic activity** *of specific departments or…*

Si la substitució per un sinònim falla, les paraules que relacionen conceptes (per exemple, preposicions) se substitueixen per un símbol especial que fa que es comptin com a coincidències totes les paraules que hi ha entre els termes relacionats. El sistema mostra els contextos de la pàgina de resultats en què els conceptes estan relacionats per una cadena de paraules, sense signes de puntuació entremig, marcada amb negreta. En aquesta cadena l'usuari hi pot veure una preposició diferent de la que ha utilitzat o pot conèixer una manera idiomàtica de relacionar les paraules que denoten els conceptes. Els contextos s'etiqueten, i es creen segments sintàctics amb la finalitat de presentar primer els contextos en què les paraules amb negreta formen el mateix segment sintàctic de la hipòtesi. Per exemple, si l'usuari escriu *we carried up a project that lasted 2 years*, on *carried up a project* és nou, el corrector primer mostra contextos com *How we* **carried out our project**, que poden ser útils perquè l'usuari s'adoni que hauria d'haver utilitzat la preposició 'out'.

S'està treballant en la possibilitat de mostrar contextos on algunes paraules que denoten conceptes i que coexisteixen en un segment sintàctic del text (sense signes de puntuació entremig)

apareguin coexistint en un segment sintàctic diferent però més freqüent. L'usuari disposaria de més maneres idiomàtiques de dir la mateixa cosa. Per exemple, mostraria **search results page** (un SN amb 515.000 resultats) en cas que l'usuari hagués escrit *the page that shows the results of the search* (1 resultat). El sistema hauria de considerar aquest SN complex com una manera més curta d'expressar les relacions conceptuals escrites a la frase.

5. Comparació amb els correctors tradicionals

El corrector que presentem és diferent dels correctors gramaticals i d'estil tradicionals, perquè no es basa en regles predefinides ni dependents de la llengua (Naber, 2003), ni en anàlisis sintàctiques (Jensen *et al.*, 1993), ni en mètodes estadístics (Atwell i Elliot, 1987). Els mòduls, excepte l'etiquetador, actuen amb un motor de cerca que no és «dependent de la llengua». Per tant, el corrector és fàcil d'adaptar a una altra llengua sempre que hi hagi un etiquetador per a aquesta llengua i el nombre de pàgines web disponibles a la Xarxa sigui elevat. Per altra banda, aquest corrector pot advertir l'usuari d'un ventall més ampli de fenòmens que superen la concordança subjecte-verb i altres errors típics tractats pels sistemes tradicionals. De fet, aquest corrector es desenvolupa com un complement d'aquests sistemes. Els correctors tradicionals ja detecten els errors d'ortografia i els errors gramaticals típics; per tant, he presentat una manera simple d'ajudar l'usuari que té mancances difícils de ser detectades per unes regles predefinides. El sistema tot just s'està desenvolupant i encara no hi ha dades per a avaluar la seva actuació. Per tant, encara no hem fet una comparació exhaustiva amb altres correctors.

6. Treball futur

En primer lloc, cal avaluar de quina manera el corrector supera alguns problemes que són inherents en les cerques per Internet. Per exemple, les pàgines web amb errors gramaticals i ortogràfics no estan discriminades i, per tant, el corrector no sap del cert si un segment que no és nou coincideix amb l'error d'un escriptor no angloparlant. Les coincidències trobades, independentment que les paraules vagin amb majúscula o minúscula, també fan que segments mal escrits no es considerin nous. Segons Naber (2003), el Google troba el segment no gramatical 'the is' perquè hi ha un document a la Xarxa que el conté: *About the IS associates*, en què 'IS' probablement és un acrònim.

Segments que no són nous, però que no són gramaticals, haurien d'aparèixer poc a la Xarxa; ara bé, quin és el nombre mínim d'aparicions per a considerar que un segment és gramaticalment correcte? Quan les paraules coexistents són molt freqüents, el llindar pot ser alt (per exemple, 'machine translation', 280.000 resultats), però la presència d'una combinació menys freqüent en un segment gramatical pot fer que el nombre de resultats caigui en picat (per exemple, 'machine translation methods', 109 resultats); per tant, el nivell s'hauria d'establir de manera coherent. Es podrien aplicar mètodes estadístics per a establir el llindar de resultats, encara que es poden considerar altres mètodes, com la identificació d'URL de confiança arran dels contextos mostrats. Per exemple, els documents provinents d'URL amb *.edu* o els documents que continguin *www.citeseer*, la gran biblioteca en línia de publicacions científiques, probablement són escrits en un anglès acceptable.

Un altre problema inherent a les cerques per Internet és la manca de criteris lingüístics dels motors a l'hora de fer les cerques. Per exemple, si es busca *I loved the woman*, el motor no compta una pàgina que contingui *I love the women*. Esperem que les consultes a Wordnet i també l'etiquetatge i segmentació dels contextos pugui atenuar aquests efectes. Aquests temes seran analitzats i quantificats en un futur pròxim.

Bibliografia

ATWELL, E.; ELLIOT, S. (1987). «Dealing with ill-formed English text». A: *The Computational analysis of English*. Longman.

JENSEN, K.; HEIDRON, G.E.; RICHARDSON, S.D. (ed.) (1993). *Natural language processing: the PLNP approach*. Kluwer Academic Publishers.

LANGKILDE, I.; KNIGHT, K. (1998). «Generation that exploits corpus-based statistical knowledge». A: *Proceedings COLING-ACL*.

LANGKILDE, I. (2002). «An Empirical Verification of Coverage and Correctness for a General-Purpose Sentence Generator». A: *Proceedings of the International Language Generation Conference 2002*. Nova York. Pàg. 17-24.

NABER, D. (2003). *A Rule-Based Style and Grammar Checker*. Universitat de Bielefeld.

SCHMID, H. (1994). «Probabilistic Part-of-Speech Tagging Using Decision Trees». A: *Proceedings of the First International Conference on New Methods in Natural Language Processing (NemLap-94)*. Manchester, Regne Unit. Pàg. 44-49.

Joaquim Moré
Investigador de l'Internet Interdisciplinary Institute (IN3) de la UOC
jmore@uoc.edu

Investigador de l'IN3 i tècnic del Servei Lingüístic de la Universitat Oberta de Catalunya especialitzat en tecnologies lingüístiques. És llicenciat en Filologia anglesa i té el màster de Lingüística computacional per la Universitat de Barcelona. Actualment desenvolupa la seva tesi doctoral entorn de l'avaluació de la traducció automatitzada.