TONI BADIA, ÀNGELS EGEA, TONI TUELLS

CATMORF, UN ANALITZADOR MORFOLÒGIC PER AL TRACTAMENT AUTOMÀTIC DE CORPUS TEXTUALS EN CATALÀ

INTRODUCCIÓ

En aquest article descriurem les característiques generals de CATMORF, el primer analitzador morfològic de dos nivells per al català, que s'utilitza en el processament de textos reals (especialitzats i periodístics principalment). En primer lloc situarem aquest projecte en relació amb les principals tendències dins la lingüística computacional, i més concretament dins la morfologia computacional. Seguidament definirem les estratègies que es poden adoptar a l'hora de construir un analitzador morfològic tenint en compte les característiques de la llengua que s'ha de tractar. A continuació presentarem el formalisme de dos nivells (considerat un estàndard dins la morfologia computacional) i explicarem la manera com s'ha aplicat al català. Finalment explicarem com s'ha dut a terme la descripció morfològica del català, pensada per ser implementada en forma de regles de dos nivells, i les solucions concretes que s'han adoptat dins d'aquest formalisme per donar compte de tota la casuística de la morfologia catalana.

LA MORFOLOGIA COMPUTACIONAL

En el tractament lingüístic, com en tants altres camps de l'activitat humana, els ordinadors han modificat els plantejaments i les possibilitats. Per mencionar només un exemple, actualment resulta molt més fàcil, ràpid i còmode escriure que no ho era fa uns anys quan el recurs tecnològic més avançat era la màquina d'escriure. Canviar parts d'un text, reorganitzar-lo, donar-li formats diferents, guardar-lo en una base de dades i accedir-hi després, corregir-lo i fins tot tra-

duir-lo són tasques que resulten molt més fàcils gràcies a diversos programes informàtics. Ara bé, entre els programes que usem en l'activitat lingüística podem fer-ne dos grans grups: els programes que manipulen el material lingüístic o textual de manera externa i els que ho fan tenint en compte les estructures lingüístiques que tenen. Els primers programes tracten el llenguatge com podrien tractar qualsevol altre element simbòlic (per exemple, dibuixos o xifres), de manera que no hi apareixen implicades nocions i estructures pròpiament lingüístiques. Els altres, en canvi, són programes que reprodueixen d'alguna manera activitats lingüístiques que fem les persones.

Des de fa força anys ja, s'ha anat consolidant el camp científic que solem conèixer sota el nom de *lingüística computacional*, que agrupa els treballs informàtics sobre el llenguatge. En gran part, la consolidació del camp ha suposat l'adopció d'objectius, mètodes i tècniques pròpies, respecte tant de la lingüística teòrica com d'altres aspectes de la intel·ligència artificial. Des del final de la dècada dels 80, la lingüística computacional ha entrat en els plans d'estudi de facultats d'informàtica i de lingüística de diversos països europeus. Paral·lelament han sorgit diversos manuals de nivell universitari que constitueixen una bona introducció al camp: ALLEN, 1995; GAZDAR i MELLISH, 1989; SAINT-DIZIER, 1991; són introduccions a la lingüística computacional que ofereixen (cada una des d'una perspectiva diferent) una visió general de les possibilitats del camp.

El tractament de la morfologia del llenguatge natural (i de les llengües particulars) és una de les qüestions que s'aborden en lingüística computacional: cada un dels tres manuals que hem citat hi dediquen un o més capítols. Ara bé, en aquesta qüestió, com en tantes altres, es poden adoptar dues perspectives radicalment diferents: o bé de cara a obtenir un millor coneixement del llenguatge, o bé de cara a construir aplicacions que resolguin algun problema pràctic. En el primer cas, ens trobarem amb plantejaments computacionals molt pròxims als de la lingüística teòrica; mentre que en el segon tendirem a buscar aquelles solucions que ens resulten més adequades de cara a la tasca que ens plantegem.

Des de la primera perspectiva, buscaríem un tractament informàtic que permeti de modelar el coneixement que tenim de l'estructura

morfològica de les llengües. En aquest sentit, intentaríem de representar de manera adequada les relacions entre les estructures fonològiques i les morfològiques, ens interessaria de partir de categoritzacions (de morfemes, afixos, etc.) que fossin vàlides des d'un punt de vista teòric, procuraríem que el nostre tractament pogués explicar tant els fenòmens regulars com els irregulars... En definitiva, procuraríem que el nostre programa reflectís els plantejaments i les solucions que avui dia la teoria morfològica proposa. I amb tota probabilitat l'adopció d'aquesta perspectiva faria replantejar alguns aspectes de la teoria morfològica, per tal com el grau de formalització al qual obliga la implementació computacional d'una teoria presentaria els problemes i les seves solucions sota una nova llum.

Si adoptéssim la segona perspectiva, en canvi, ens interessaria de posar a punt un programa que fes les operacions morfològiques d'anàlisi i generació d'una manera eficient i efectiva enfront de textos reals. Un programa així no posaria l'accent en els plantejaments teòrics, sinó en les operacions pràctiques que el programa hauria de dur a terme; és a dir, fer l'anàlisi morfològica de textos diversos i generar la forma superficial pertinent a partir d'estructures informatives sobre les propietats morfològiques d'un mot. En aquesta perspectiva, els aspectes pràctics poden fer prendre decisions que des d'un punt de vista teòric potser no estan del tot justificades.

Naturalment els dos plantejaments són legítims i poden coexistir sense problemes en el si de la disciplina que hem anomenat *lingüística computacional*. La dualitat de les dues perspectives no és més que la concreció en aquesta disciplina de la dualitat entre teoria i aplicació, que és comuna a tots els àmbits del saber (especialment els més influïts pel progrés tecnològic). No obstant això, en els darrers anys s'ha anat imposant una diferenciació terminològica per a distingir entre les dues perspectives: s'ha tendit a reservar la denominació de *lingüística computacional* per a l'orientació més teòrica, mentre que s'ha creat el terme d'*enginyeria lingüística* per a la perspectiva més aplicada. Sens dubte, en l'èxit d'aquesta denominació hi han contribuït diversos factors, dels quals aquí només enumerarem els que considerem més decisius: la tendència excessivament teòrica dels treballs en el camp durant els anys 70 i 80, l'adopció de mètodes de la lingüística

computacional per part d'investigadors que fins ara només s'havien mirat els textos com a símbols per a ser manipulats no lingüísticament, l'augment extraordinari de les prestacions dels ordinadors (capacitat de memòria, temps de processament...), la pressió dels organismes finançadors públics (i d'una manera especial la Unió Europea amb els seus programes sobre les tecnologies de la informació), i el major grau de popularitat d'un terme com *enginyeria*, enfront del de *lingüística*.

L'analitzador morfològic que presentem en aquest article adopta clarament la segona perspectiva. Es tracta d'un analitzador morfològic pensat per a l'anàlisi morfològica de textos reals. El tractament de corpus textuals és justament una de les tasques en què la lingüística computacional aplicada (o, si ho preferiu, l'enginyeria lingüística) permet de dur a terme tasques que altrament seria pràcticament impossible de realitzar.

L'estat actual de desenvolupament tecnològic en enginyeria lingüística ens permet obtenir un nivell d'encert equivalent a l'humà en moltes de les tasques a realitzar en anàlisi textual (per exemple, l'anàlisi de freqüències o l'anàlisi i desambiguació morfològiques d'una paraula) i inferior en d'altres (per exemple, l'anàlisi sintàctica profunda de frases). La tecnologia informàtica aplicada al tractament automàtic de corpus textuals ens permet, avui en dia, uns resultats més que acceptables en les següents operacions bàsiques:

— detecció i segmentació d'unitats textuals bàsiques (paràgrafs, frases, etc.);
— detecció de noms propis, dates, números, unitats fraseològiques;
— anàlisi morfològica de paraules simples; assignació a aquestes de lema i informació morfològica simple (gènere, nombre, persona, mode, etc.);
— desambiguació morfològica de paraules morfològicament ambigües;
— anàlisi sintàctica superficial de frases (detecció d'alguns tipus de sintagmes, etc.)

En aquest article ens centrem en l'anàlisi morfològica. El disseny i implementació d'un analitzador morfològic per a qualsevol llengua obliga a estudiar els aspectes següents:

1) mecanismes de formació de paraules de la llengua a tractar (per exemple, l'anglès és una llengua morfològicament «pobra», en l'alemany es fa servir la composició més sovint que en català, i el basc és una llengua aglutinant);
2) motivació teòrica que hi hagi al darrera de la nostra implementació (algunes solucions que es proposen no tenen cap mecanisme explicatiu o predictiu de la morfologia de les paraules, mentre que d'altres sí que tenen en compte aquests aspectes);
3) eficiència de la solució proposada (rapidesa, espai,...) (solucions teòricament interessants poden resultar en un programa terriblement lent, i a la inversa);
4) facilitat d'ampliació i modificació del sistema (incorporació de noves paraules o fenòmens morfològics a tractar).

D'aquestes consideracions generals se'n deriva la conclusió que no hi ha una única manera de fer un analitzador morfològic: per exemple, una estratègia raonable per a l'anglès pot ser inviable per al basc. De fet, aquests quatre aspectes que acabem d'enumerar poden constituir altres tants paràmetres o criteris amb què avaluar les propostes d'implementació de l'anàlisi morfològica. Així doncs, els tindrem en compte en la propera secció, quan analitzem les diferents propostes bàsiques de tractament computacional de la morfologia.

A mode d'introducció general, podem afirmar que l'analitzador que presentem, CATMORF (CATMORF, 1997), és usat en l'anàlisi morfològica de textos reals: en aquests moment, s'hi processen tant els textos catalans del corpus especialitzat de l'Institut Universitari de Lingüística Aplicada com textos de diari («Avui» i «El Periódico»). Ara bé, en la nostra implementació hem intentat de reproduir en la mesura que ha estat possible l'estat de la qüestió en anàlisi morfològica computacional; és a dir, hem buscat de realitzar un programa pràctic, però alhora satisfactori des d'un punt de vista teòric (en el marc de la lingüística computacional). Hem buscat, doncs, un cert punt intermedi entre la

mera implementació de la teoria lingüística i la pràctica més absoluta. Les raons per adoptar un plantejament així són diverses. En primer lloc, resulta més satisfactori des d'un punt de vista científic la realització d'un programa que suposi algun repte. En segon lloc, és clar que els plantejaments de programació més actuals (en aquest cas, en morfologia computacional) tenen en compte l'estat actual de les tècniques de *hardware* i de *software*; en aquest sentit cal tenir en compte que programar l'any 1998 amb tècniques, per exemple, de l'any 1995 permetrà un millor aprofitament que si haguéssim programat amb tècniques pròpies de l'any 1980. En tercer lloc, hem partit d'una concepció totalment modular de la programació, de manera que el problema general de l'anàlisi morfològica és descompost en problemes més petits (p. ex., la segmentació dels mots i la construcció de l'estructura de la paraula, tot diferenciant entre categories nominals i categories verbals, etc.); a la vegada, hem procurat que el nostre programa s'integri adequadament (com un mòdul entre altres) amb altres programes del seu entorn (gestor de la grafia i forma de les frases, gestor del lèxic, programa de desambiguació morfològica, etc.). En quart lloc, hem procurat que els usuaris del programa (normalment lingüistes) puguin usar-lo sense massa dificultats; això té repercussions directes sobre la forma com es codifica la informació: com més explícita sigui la informació lingüística que reclama el programa més fàcil serà que els seus usuaris la tractin bé. Finalment, ens ha interessat que el nucli del programa pugui ser utilitzat tant en mode d'anàlisi com de generació, per tal de no haver de tenir dos programes diferents per a cada una d'aquestes dues funcions, que farien servir estructures lingüístiques idèntiques.

IMPLEMENTACIÓ DE CATMORF

Abans d'entrar en detalls sobre la implementació del nostre analitzador, explicarem breument algunes estratègies bàsiques d'implementació d'analitzadors morfològics (una referència obligada per a estudiar aquestes estratègies es pot trobar a [SPROAT, 1992]). També mostrarem com s'han d'avaluar aquests procediments tenint en compte els criteris 1-4 que s'han explicat a la secció anterior.

POSSIBLES SOLUCIONS

Solució 1: Llistes de paraules

Aquesta és la solució més simple; el programa manté una llista de paraules amb la seva informació morfològica associada. Per exemple, un analitzador morfològic per al català que segueixi aquesta estratègia podria tenir una llista amb les entrades següents:

Paraula	***Morfologia***
cotxe	**cotxe**: NMS
cotxa	**cotxa**: NFS
cotxes	**cotxe**: NMP **cotxa**: NFP

A la primera columna trobem les formes que volem analitzar; a la segona trobem l'anàlisi morfològica de les formes. Per exemple, la forma *cotxe* s'analitzaria com a **cotxe:** NMS, on s'interpreta que la forma analitzada correspon al lema substantiu **cotxe** i és masculina i singular (NMS). La forma *cotxes*, en canvi, presenta dues anàlisis possibles:

cotxe (NMP)
cotxa (NFP)

Reconèixer i assignar informació morfològica a una paraula és trivial: si la paraula és a la llista se li assigna la informació del segon camp; si la paraula no és a la llista, el programa no reconeix la paraula.

L'avaluació d'aquesta proposta la farem seguint els criteris 1-4 descrits anteriorment:

1) En el cas del català, aquesta solució implica entrar totes les formes flexionades dels lemes que volem tractar (fins a 4 en el cas dels noms i els adjectius, però fins a quaranta en el cas dels verbs). Per a l'anglès aquesta solució és bastant òptima, ja que és una llengua amb poca morfologia. En el cas de l'alemany la

solució és molt poc satisfactòria, ja que la formació de paraules per composició és un fenomen molt freqüent.
2) No hi ha cap motivació teòrica al darrere d'aquesta proposta: ni mecanismes explicatius ni predictius.
3) La proposta és fàcil d'implementar i s'aconsegueixen analitzadors molt ràpids.
4) Afegir noves paraules és costós, perquè s'han d'introduir totes les formes.

Solució 2: Segmentació i combinació d'unitats simples

Aquesta solució no és pròpiament una solució, sinó un conjunt de solucions. El reconeixement d'una paraula es fa seguint els passos següents:

— segmentació de la paraula en unitats simples (prefixos, lexemes, sufixos, etc.). Aquesta és la part *morfografèmica* del procés de reconeixement.
— combinació de les unitats simples per obtenir la informació morfològica necessària. Aquesta és la part *morfotàctica* del procés de reconeixement.

Per exemple, l'anàlisi de la paraula *cases* (com a nom) es faria seguint els passos següents:

Morfografèmia: *cases* → *casa s*
Morfotàctica: lexema + sufix plural → paraula plural

En el primer pas obtindríem dues unitats simples: el lexema *casa* i el sufix de nombre *s*. En el següent pas combinem tots dos elements i deduïm que tenim una paraula en plural. Fins aquí hem explicat *què han de fer* les parts morfografèmica i morfotàctica de l'analitzador, però encara no hem dit *com s'han de dur a la pràctica* aquests processos.

Les solucions que s'han proposat en els diversos projectes diferei-

xen en la consideració de què són unitats simples, en els mecanismes de segmentació d'aquestes unitats i en la manera de combinar-les per obtenir-ne la informació morfològica.

L'avaluació d'aquesta proposta la farem seguint els criteris 1-4 descrits anteriorment:

1) No és fàcil de determinar, per a cada llengua, quines són les unitats simples a partir de les quals formem les paraules.
2) Una solució d'aquest tipus relaciona una forma amb les seves unitats simples (mecanisme explicatiu) i permet predir formes a partir de lemes (mecanisme predictiu). Això necessàriament ha d'incloure una classificació adequada de les paraules en models morfològics de comportament similar (noms, classificació dels verbs segons uns models, etc.).
3) Normalment, aquestes solucions són més lentes que les proposades a la solució 1 (són, per tant, pitjors des del punt de vista de la rapidesa en l'obtenció de resultats).
4) La incorporació de lemes nous acostuma a ser més simple. El problema principal és la correcta assignació de la paraula als models ja existents.

L'estratègia concreta que segueix CATMORF per determinar les unitats simples, la manera d'obtenir-les i la manera de combinar-les es basa en l'anomenat *paradigma de dos nivells*.

El paradigma de dos nivells

En morfologia computacional, actualment, el paradigma de dos nivells, que trobem definit a KOSKENNIEMI (1984) i KAPLAN i KAY (1994), està considerat com un estàndard, és a dir, és la tècnica de propòsit general més àmpliament acceptada i utilitzada en la descripció morfològica de diferents llengües: finès, basc, LSGRAM,[1] etc.

1. El projecte LSGRAM (LRE 61029) va consistir en l'elaboració de diferents gramàtiques computacionals de cobertura àmplia per a nou de les llengües oficials de

Morfografèmia

Per poder descriure correctament la part morfografèmica d'un analitzador morfològic de dos nivells cal introduir els conceptes de *nivell superficial* i *nivell lèxic*. El nivell superficial és la representació de la forma ortogràfica de la paraula que s'ha d'analitzar; i el nivell lèxic és la combinació de morfemes simples que hem obtingut. Per exemple:

textos (nivell superficial)
text + *s* (nivell lèxic)

textos és la forma que volem analitzar (nivell superficial), i la combinació vàlida de morfemes resultant és: ***text*** (lema) i ***s*** (sufix de nombre). El caràcter + indica la combinació de morfemes. Naturalment, hi ha un marge molt ampli de maniobra en l'elecció dels morfemes del nivell lèxic; per exemple, podríem considerar que al nostre sistema hi ha un únic sufix de nombre (*s*) o que n'hi ha 3 (*os, s, es*). Vegem-ne uns exemples:

textos
text + *s* o bé *text* + *os*

cases
casa + *s* o bé *casa* + *es*

espessos
espès + *s* o bé *espess* + *os*

---

la Unió Europea (entre les quals hi havia el castellà). El tractament morfològic d'aquestes llengües es va fer amb un formalisme de dos nivells. Es pot trobar més informació sobre el projecte LSGRAM en aquesta URL: http://www.iai.uni-sb.de/LSGRAM/home.html.

2. Les regles de dos nivells estan basades en regles fonològiques. Veg. KAPLAN I KAY, 1994.

El pas del nivell superficial al lèxic, és a dir, la part morfografèmica del sistema, es fa, doncs, mitjançant regles de dos nivells[2] (RDN), que relacionen contextos superficials amb contextos lèxics (o dit d'una altra manera, les RDN reescriuen elements del nivell superficial en elements del nivell lèxic). A CATMORF el format d'aquestes regles és el següent:[3]

regla nom-de-la-regla:
{CSE CTNS CSD operador-direcció CLE CTNL CLD}

La interpretació de les sigles utilitzades és:

CLD: context lèxic dret
CLE: context lèxic esquerre
CSD: context superficial dret
CSE: context superficial esquerre
CTNS: caràcters del nivell superficial que volem transformar
CTNL: caràcters del nivell lèxic que han estat transformats

Els contextos superficial i lèxic mostren els caràcters vàlids als quals es pot aplicar la regla.

L'operador de direcció pot ser algun d'aquests:

⇒ Els caràcters del nivell superficial s'han de convertir obligatòriament als caràcters del nivell lèxic.

⇐ Els caràcters del nivell lèxic s'han de convertir obligatòriament als caràcters del nivell superficial.

⇔ ⇒ i ⇐

opt Els caràcters del nivell superficial opcionalment es poden convertir als caràcters del nivell lèxic, i viceversa.

3. Vegeu SEGMORF, 1996 i SEGMORF, 1997 per a una descripció detallada d'aquest format.

Per exemple, la regla següent ens converteix (o reescriu) la *o* superficial de *textos* en un caràcter de separació en el nivell lèxic:

regla *o*_epent: {[x, t] [o] [s] ⇒ [x, t] [+] [s]}

Interpretació de la regla: el caràcter *o* del nivell superficial s'ha de convertir en el caràcter + del nivell lèxic si va precedida dels caràcters *xt* i seguida de *s*. L'efecte pràctic d'aquesta regla és que eliminem la *o* quan analitzem i introduïm el caràcter de separació entre morfemes (per exemple, *textos)*.

L'exemple de *cases* es podria resoldre amb la regla següent:

regla canvi_es: {[ ] [e, s] [ ] ⇐ [ ] [a, +, s] [ ]}

Interpretació de la regla: si una paraula acaba en *a* al nivell lèxic, i li apliquem flexió de nombre plural (a+s), ha d'aparèixer *es* en el nivell superficial.

Morfotàctica

La part morfotàctica d'un analitzador morfològic de dos nivells consisteix a assignar una informació morfològica a la combinació de morfemes (obtinguda de les RDN) mitjançant *regles de formació de paraules*. El conjunt de regles de formació de paraules necessari per donar compte d'un determinat fenomen constitueix la *gramàtica de paraula* (GP). Per exemple, per al reconeixement de *textos* tindríem una regla de formació de paraules com aquesta:

lema (masc, singular) + sufix (nombre) → paraula (masc, plural)

Interpretació de la regla: si un lema masculí i singular es combina amb un sufix de nombre, hem reconegut una paraula que té gènere masculí i nombre plural.

Per al reconeixement de *cases* podríem tenir una regla com:

lema (fem, singular) + sufix (nombre) → paraula (fem, plural)

Interpretació de la regla: si un lema femení singular es combina amb un sufix de nombre, hem reconegut una paraula que té gènere femení i nombre plural.

Per al reconeixement de la forma *intel·ligents*:

lema (masc-fem, singular) + sufix (nombre) → paraula (masc-fem, plural)

Interpretació de la regla: si un lema masculí/femení i singular es combina amb un sufix de nombre, hem reconegut una paraula que té gènere masculí/femení i nombre plural. Fixem-nos, però, que aquesta regla no tractaria adequadament casos com *capaç* i *atroç*, en què la forma singular és de gènere invariable, però la forma plural és de gènere variable (*capaços, capaces*).

Per al reconeixement de les formes verbals podem seguir una estratègia similar i escriure regles com aquestes:

Morfografèmia:	*canto* → *cant* + *o*
Morfotàctica:	arrel (1a) + sufix (1-persona, present) → forma (1-persona, present)

Aquests exemples són senzills, però a la pràctica ens trobem que la casuística és força més complicada, sobretot si hi afegim regles per tractar processos derivatius.

Les regles de dos nivells i la gramàtica de paraula de CATMORF

Així doncs, un analitzador morfològic de dos nivells està format per un conjunt de regles de dos nivells, un conjunt de regles de formació de paraules i un programa que manipula tots dos conjunts per assignar informació morfològica a la paraula que analitzem.

Pel que s'ha vist als exemples anteriors, tenim diferents graus de llibertat per escollir:

— quins són els elements simples del sistema (per ex., sufixos *s*, *os*, *es*);
— quines són les RDN (segons la selecció d'elements simples feta anteriorment);
— quines són les regles de la GP a partir dels elements escollits als apartats anteriors

Respecte a la flexió nominal, hem optat per tenir un nombre com més reduït millor d'elements simples: hi ha un únic sufix de nombre (*s*) i un únic sufix de gènere (*a*). Això ens ha permès de tenir un nombre més reduït de regles de la GP, i un nombre més elevat de RDN que si haguéssim optat per un nombre més gran d'elements simples (sufixos de nombre: *s*, *os*, *es*, etc.). La motivació principal d'aquesta elecció ha estat el d'*elegància teòrica*.

Respecte a la flexió verbal, hem optat per analitzar les formes verbals com a combinació d'una arrel i un sufix (per exemple, *cantava* s'analitza com a **cant** + **ava**). Això ens ha permès de reagrupar sufixos comuns a diversos models i simplificar extraordinàriament les regles de formació de paraules. (El fet que la GP sigui una gramàtica d'unificació permet que només hi hagi una única regla de formació de verbs).

DESCRIPCIÓ MORFOLÒGICA DEL CATALÀ

Les decisions preses en la construcció d'un analitzador morfològic de dos nivells depenen fonamentalment dels resultats observats en un estudi previ que posi de manifest tota la casuística relacionada amb els fenòmens morfològics de què es vol donar compte, que en aquest cas són la flexió nominal i la flexió verbal. Aquest estudi, al seu torn, té l'objectiu d'observar els canvis sistemàtics que es produeixen entre els dos nivells i determinar els elements que permeten expressar aquests canvis en forma de generalitzacions. Aquestes generalitzacions, que posteriorment s'implementaran en forma de regles, han de basar-se en les informacions de què disposarà l'analitzador.

Aquestes informacions són els lemes del nivell lèxic de CATMORF, que s'han obtingut automàticament del diccionari de l'Institut d'Estu-

dis Catalans (DIEC)[4] (i posteriorment s'ha ampliat amb el diccionari d'Enciclopèdia Catalana [DEC83]). El lema és la representació abstracta d'un mot, que consta de categoria gramatical i una forma d'identificació que convencionalment es fa coincidir (com en les entrades de diccionari) amb la forma morfològicament menys marcada: singular per als substantius, masculí per als adjectius i infinitiu por als verbs:

Forma	**Lema** (nivell lèxic)
fosques	**fosc** adj
llibres	**llibre** m
convenen	**convenir** v intr

Això fa innecessari tractar les formes flexives corresponents al singular i al masculí, ja que es troben implícites en la mateixa representació del lema. I la resta de formes flexives cal estudiar-les amb relació al lema. Per això, els elements que s'han pres en consideració com a condicionats de la flexió són la categoria gramatical i les característiques gràfiques del lema. I en aquests dos paràmetres s'han basat totes les correspondències entre els dos nivells.

La categoria gramatical permet determinar el tipus de flexió d'un mot: un substantiu d'un sol gènere només podrà tenir flexió de nombre; un adjectiu, només de gènere i nombre; un verb, només de persona, temps i mode. Les característiques gràfiques de cada mot permeten determinar, dins de cada tipus de flexió, la classe flexiva a què pertany, és a dir, les marques flexives que se li poden assignar i les conseqüències gràfiques que l'assignació d'aquestes marques pot comportar (aparició o desaparició d'accents, canvi d'accent per dièresi, etc.):

Lema	**Marca de plural**	**Forma**
forat m	*-s*	*forats*
patí m	*-ns* + supressió d'accent	*patins*
veí m	*-ns* + canvi d'accent per dièresi	*veïns*
fosc adj	*-os*	*foscos*

4. El procés automàtic s'explica a (TUELLS, 1998).

El resultat d'aquest estudi és una classificació dels mots en classes flexives a partir de la categoria i la forma del lema, i en una classificació de les marques flexives segons la seva naturalesa (marques de plural, de femení, etc.) i el tipus de lema a què s'adjunten. La descripció detallada de les condicions que distingeixen els mots agrupats en una mateixa classe flexiva per oposició a les condicions descrites per a la resta de les classes és tan important com la classificació mateixa, per tal com aquestes condicions són la base per a la formulació de les regles. Per exemple:

Marca de plural: *-os*
Condicions
Caràcters finals del lema: ***-sc***
Categoria gramatical de lema: m, adj
Ex: *fosc foscos*

Els segments definits com a marques flexives no pretenen ser una proposta teòricament fonamentada de segmentació dels mots, sinó una estratègia vàlida per a la classificació dels fenòmens estudiats, tenint en compte que la unitat de partida és el lema, que no deixa de correspondre gràficament a una forma flexionada. Per tant, les marques flexives indicades representen els segments canviants entre el nivell lèxic i el nivell superficial.

Abast de la descripció morfològica

En aquesta primera fase del projecte només s'ha tractat de la flexió nominal de gènere i nombre, i de la flexió verbal de temps, mode i persona. No s'han tingut en compte les realitzacions gràfiques normativament incorrectes més habituals, amb què l'analitzador es podria topar en ocasions, ni les formes morfològiques dialectals, que són molt abundants en la flexió verbal.

Com a corpus de treball s'han estudiat totes les entrades lèxiques del DIEC. Això vol dir que en aquests moments les regles només donen compte de les formes per a les quals hi ha un lema incor-

porat en el diccionari; les formes sense lema previst al diccionari no es resolen.

En una etapa posterior està previst de completar aquest estudi amb la descripció i posterior implementació de dos aspectes importants que milloraran els resultats obtinguts en aquest moment:

— Les formes aspectives dels substantius i adjectius (diminutius, augmentatius, etc.).
— Els sufixos derivatius. L'objectiu d'aquest estudi és la determinació dels elements necessaris per a la predicció de lemes sense entrada en el diccionari.

En aquesta primera fase d'anàlisi no es té en compte el context i, per tant, no és possible de desambiguar ocurrències coincidents amb formes idèntiques de lemes diferents. Per exemple:

creient	**creient** m
creient	**creure** v tr

Això vol dir que per a cada ocurrència *creient*, l'analitzador donarà dues lectures possibles.

Marques flexives i classes flexives

L'estudi de la flexió s'ha dividit en dos grans blocs: morfologia nominal i morfologia verbal. La morfologia nominal tracta de la flexió dels substantius i els adjectius conjuntament, però tracta independentment la flexió de gènere i la flexió de nombre. La morfologia verbal, en canvi, tracta conjuntament la flexió de persona, temps i mode.

Això vol dir que no s'han definit marques que permetin identificar al mateix temps segments que combinin gènere i nombre. Segons la nostra anàlisi, una seqüència com *distributives*, que correspon a la forma femenina plural de l'adjectiu *distributiu*, conté dues marques, una de femení (*-a* + canvi de *-u* per *-v*) i una de plural (*-s* + canvi de *-a* per *-e*).

En la descripció de la flexió verbal s'ha optat per mantenir en un sol segment totes les informacions morfològiques relatives a persona, temps i mode. Així, la forma *cantaríem* admet una única segmentació: *cant -aríem*, on el segment *-aríem* informa que es tracta de la primera persona del plural del temps condicional, mode indicatiu.

La flexió nominal i verbal del català s'ha dut a terme, doncs, tenint en compte únicament tres tipus o grups de marques flexives:

marques de plural
marques de femení
marques de conjugació verbal

L'existència d'una marca flexiva comporta l'existència d'una classe flexiva de mots. Hi haurà, per tant, tantes classes flexives com marques diferents. Els mots que tenen flexió de gènere i nombre pertanyen a dues classes flexives, ja que s'ha estudiat per separat la flexió de gènere i la flexió de nombre.

D'altra banda, l'existència d'una classe flexiva està condicionada per la possibilitat de delimitar un grup de mots amb unes característiques flexives comunes previsibles a partir dels dos paràmetres abans esmentats: categoria gramatical i grafia del lema. L'establiment de classes flexives, per tant, està subjecta a la possibilitat de definir unes condicions distintives per a cada classe.

Moltes de les classes flexives establertes contenen un nombre més o menys important d'excepcions, és a dir, de mots que tenen una altra marca flexiva però que no presenten cap especificitat en la forma i categoria del lema que permeti distingir-los com a classe. Per exemple, els lemes adjectius acabats en *-ant* o *-ent* són invariables de gènere. No obstant això, hi ha un nombre d'adjectius amb la mateixa terminació que fan el femení en *-a* final. Com que no hi ha cap element que ens permeti preveure quins són variables i quins no ho són, no queda altra alternativa que tractar aquests darrers com a excepcions de la regla, ja que en són un nombre menor.

FLEXIÓ NOMINAL

Formació del plural

Les marques de formació del plural s'apliquen a tots els mots amb categoria gramatical d'adjectiu o substantiu. La formació del plural en català consta de 5 marques flexives si tenim en compte únicament el segment que s'afegeix al lema. Però cal considerar que existeixen almenys 14 classes flexives de mots si tenim en compte els canvis gràfics que provoca en el lema l'addició d'aquestes marques.

1) Marca de plural **zero** (en què el lema i la forma de plural coincideixen): *pelvis*
2) Marca de plural ***-s***
   a) sense modificacions en la forma del lema: *torre torres*
   b) amb canvi de *-a* per *-e*: *saba sabes*
   c) amb aparició d'accent greu (llevat de *-i*) a la penúltima vocal: *col·lagen col·làgens*
3) Marca de plural ***-ns***
   a) sense modificacions en la forma del lema: *fi fins*
   b) amb supressió de l'accent final del lema: *camió camions*
   c) amb canvi d'accent per dièresi: *veí veïns*
4) Marca de plural ***-os***
   a) sense modificacions en la forma del lema: *fosc foscos*
   b) amb supressió d'accent: *abundós abundosos*
   c) amb supressió d'accent i duplicació de la s final del lema: *espès espessos*
   d) amb canvi d'accent per dièresi: *país països*
   e) amb canvi de *-ig* per *-j*: *raig rajos*
   f) amb canvi de *-ig* per *-tj*: *lleig lletjos*
5) Marca de plural ***-es***
   a) amb canvi de *-ç* per *-c*: *feliç felices*

La darrera marca de plural només s'aplica a formes femenines d'adjectius. En la resta dels casos no es fa referència a la formació del femení plural, ja que es troba implícita en la classe flexiva 2b, que correspon a les formes singulars (substantives o adjectives) acabades en *-a*.

Cada marca flexiva s'associa a un grup de condicions o restriccions d'aplicabilitat de tipus gramatical i gràfic. Per exemple, la marca de plural zero, que implica que la forma del lema i la forma de plural són idèntiques, és pròpia dels mots el lema dels quals reuneix les característiques següents:

Caràcters finals del lema: vocal + *-s*
Categoria gramatical: m, adj
Altres condicions: mot no agut i no monosíl·lab

o bé

Caràcters finals del lema: *-s*
Categoria gramatical: f

Aquestes condicions permeten la generació o reconeixement de formes plurals com: *penis, focus, pelvis*, etc. Però eviten que es considerin plurals seqüències com: *nas, tos, mes, pedrís*, etc.

Formació del femení

Les marques de formació del femení s'apliquen a les formes singulars dels adjectius i substantius de doble gènere. Per a la formació d'un adjectiu o substantiu femení plural, cal primer formar el femení i després aplicar les marques de plural.

D'acord amb el nostre estudi, la formació del femení en català consta de 4 marques de femení i 14 classes flexives. Encara que el nombre final de classes flexives de femení coincideixi amb el nombre de classes flexives per al plural, el contingut de les classes no és coincident.

1) Marca de femení **zero** (en què el lema i la forma de femení coincideixen): *indígena*
2) Marca de femení ***-a*** *afegida* al lema (o forma masculina)
   a) sense modificacions en la forma del lema: *amorf amorfa*
   b) amb aparició d'accent greu (excepte *í* i *ú*) a la penúltima vocal: *aeri aèria*

   c) amb canvi de *-u* per *-v*: *viu viva*
   d) amb canvi de *-òleg* per *-òlog*: *filòleg filòloga*
   e) amb canvi de *-ig* per *-j*: *boig boja*
   f) amb supressió d'accent: *ofès ofesa*
   g) amb supressió d'accent i duplicació de *s*: *espès espessa*
   h) amb canvi de *-t* per *-d*: *acabat acabada*
3) Marca de femení ***-a*** *substituint* la darrera vocal del lema: *alumne alumna*
4) Marca de femení ***-na***
   a) sense modificacions en la forma del lema: *fi fina*
   b) amb supressió de l'accent: *dejú dejuna*
   c) amb canvi d'accent per dièresi: *roí roïna*

Les regles de formació del femení tenen essencialment la mateixa estructura que les regles de formació de plural, si bé presenten en general un grau de complexitat superior, així com un nombre molt elevat d'excepcions.

Implementació

Un total de 114 regles de dos nivells i 15 regles de gramàtica de la paraula donen compte de tota la flexió nominal. Alguns exemples de regles de dos nivells són:

regla afegir_n1: {[i,o] [n] [s] ⇐ [i,ó] [+] [s]}
regla afegir_n2: {[e,ï] [n] [s] ⇐ [e,í] [+] [s]}
regla u_per_v: {[Vocal] [v] [ ] ⇐ [Vocal] [u] [+,a]}
regla leg_log: {[Vocal-accentuada] [l,o,g] [ ] ⇐ [Vocal-accentuada] [l, e, g] [+,a]}

Les 15 regles de gramàtica de la paraula permeten explicar tota la combinatòria possible entre arrels substantives i arrels adjectives, i sufixos de gènere i de nombre (arrels substantives de gènere invariable i nombre variable, arrels adjectives de gènere invariable quan són singulars però de gènere variable quan són plurals, etc.).

Flexió verbal

La descripció morfològica dels verbs catalans parteix de la clàssica divisió en tres conjugacions, que permet agrupar els verbs per la terminació d'infinitiu i, alhora, per les similituds en el paradigma flexiu entre els verbs amb una mateixa terminació. Cadascuna d'aquestes conjugacions, al seu torn, s'ha subdividit en diversos grups segons la diversitat de paradigmes verbals diferents que s'han observat.

Per a aquesta classificació s'ha partit d'una llista exhaustiva de tots els verbs que apareixen al DIEC. Això assegura, per una banda, que la descripció prevegi totes les possibilitats flexives dels verbs catalans i, per una altra, la inclusió de totes les decisions preses darrerament per la Secció Filològica que afecten la flexió d'alguns verbs.

Per a l'establiment de les classes flexives verbals, s'ha partit de la forma de l'infinitiu, que és també la forma que pren el lema de cada verb. Aquesta classificació és especialment rellevant en els casos en què la forma gràfica de l'infinitiu determina el comportament flexiu del verb d'una manera previsible.

Per establir les correspondències entre els lemes i les seves realitzacions flexives, s'ha partit del supòsit que tot verb consta d'una part generalment invariable, que és pròpia de cada verb, i una part variable o flexiva, que és pròpia d'un conjunt més o menys ampli de verbs. Direm que tenen el mateix paradigma verbal els verbs que presenten una flexió idèntica; és a dir, els verbs que comparteixen la part o segment variable en cadascuna de les seves realitzacions. Cadascun d'aquests conjunts de verbs constitueix un grup verbal.

Cal aclarir que la distinció entre segment invariable i segment variable d'un verb no coincideix necessàriament amb la divisió entre arrel i desinències. En cada cas s'ha fet la segmentació que s'ha considerat més oportuna per tal de representar amb la màxima simplicitat possible la flexió d'un grup determinat de verbs. En general, s'ha considerat invariable aquella part del verb que es manté inalterable en tota la flexió.

Això vol dir que els segments variables poden tenir extensions diferents d'un grup a un altre encara que pertanyin a la mateixa conjugació. Per tant, és imprescindible indicar per a cada classe flexiva quin és el segment variable i, per tant, substituïble. Com que s'ha convin-

gut de donar al lema la mateixa forma que té l'infinitiu, s'indica per a cada grup verbal quin és el segment variable de l'infinitiu, que, com a forma flexiva que és, haurà de ser compartit per tot el grup.

Primera conjugació

La primera conjugació presenta el paradigma més regular i previsible de tota la morfologia verbal, tant per als verbs existents en el diccionari com per als verbs que amb el temps s'hi vagin incorporant. (Afortunadament també és la més productiva.) Això ha permès fer-ne un tractament diferenciat respecte a la resta de conjugacions, que són molt poc productives i presenten un grau molt elevat d'irregularitats que fa difícil de trobar models clars de conjugació.

Llevat d'un nombre molt reduït de verbs que es poden considerar excepcions, es poden trobar deu paradigmes regulars diferents. Aquests deu paradigmes, però, es redueixen a un paradigma bàsic i nou de complementaris deduïbles a partir del primer i a partir de les característiques gràfiques de l'infinitiu, concretament dels caràcters immediatament precedents al segment final *-ar* de l'infinitiu. En gairebé tots els casos es tracta de canvis de naturalesa ortogràfica, com és el cas de l'alternança de les grafies *g/gu, ç/c, j/g*, etc.

Atesa la regularitat i previsibilitat d'aquestes alternances purament ortogràfiques, s'ha preferit tractar el segment *-ar* com a únic segment variable per a tota la conjugació, de manera que hi ha un sol paradigma verbal amb alternances gràfiques en el segment invariable. Per a cada tipus d'alternança s'ha definit un grup complementari del paradigma bàsic, que no pateix cap alteració en el segment no substituïble.

Així, doncs, per a la primera conjugació hi ha un sol grup verbal amb el quadre de conjugació complet, representat pel verb *cantar*:

Grup 1: Infinitiu acabat en consonant + *-ar* (ex. *cantar*)

i vuit grups complementaris del primer, segons les especificitats gràfiques del segment no substituïble de cada verb:

Grup 1.1: Infinitiu acabat en *-gar* (ex. *pregar*)
Grup 1.2: Infinitiu acabat en *-car* (ex. *trencar*)
Grup 1.3: Infinitiu acabat en *-guar* (ex. *enaiguar*)
Grup 1.4: Infinitiu acabat en *-quar* (ex. *obliquar*)
Grup 1.5: Infinitiu acabat en *-jar* (ex. *envejar*)
Grup 1.6: Infinitiu acabat en *-çar* (ex. *començar*)
Grup 1.7: Infinitiu acabat en *-ear*, *-iar*, *-ouar* o *-uar* (llevat dels en *-guar* i *-quar*) (ex. *menysprear*)
Grup 1.8: Infinitiu acabat en vocal + *-iar* (ex. *desmaiar*)

Per a cadascun d'aquests grups complementaris s'han enumerat les formes en què hi ha diferències respecte al grup 1. En l'exemple següent podem veure marcades en negreta les formes del grup 1.7 no coincidents amb les del grup principal:

Diferències respecte al grup 1:

pr. Subj.	**ï**	**ïs**	**ï**	em	eu	**ïn**

A més del model bàsic de conjugació i els vuit models complementaris, hi ha sis verbs tractats independentment, bé perquè són de flexió irregular —*estar, anar, matar* (aquest darrer és irregular en una de les formes del participi)—, bé perquè no tenen una flexió completa (*dar*), o bé perquè presenten certes peculiaritats ortogràfiques (*donar*, *aguar*).

Segona i tercera conjugacions

Quant a la segona i tercera conjugacions, els grups verbals establerts no responen a l'existència d'uns paradigmes previsibles a partir de la forma gràfica del lema, sinó a l'intent de fer una descripció de tots els paradigmes verbals d'aquestes conjugacions de la manera més simple i sistemàtica possible per facilitar el tractament informatitzat posterior.

Pel fet de ser conjugacions molt poc productives en català, es fa difícil de parlar de verbs regulars, si bé és possible establir uns models

bàsics de conjugació a partir dels grups de verbs amb una terminació d'infinitiu que comparteixen un mateix paradigma o en els quals es produeix un mateix tipus de canvis gràfics.

La inexistència d'un paradigma regular per als verbs no compresos en la primera conjugació i la gran diversitat de paradigmes flexius que presenten fa possible agrupar-los de maneres molt diverses segons el criteri i elements que es tinguin en compte. El tractament que n'hem fet nosaltres és només un dels possibles. A l'hora d'establir els grups s'han tingut en compte en primer lloc la terminació de l'infinitiu, ja que aquesta és la forma del lema.

Com es pot observar, i a diferència del que hem vist en la primera conjugació i en la flexió nominal, les característiques morfològiques dels verbs d'aquestes conjugacions no han fet possible trobar uns elements descriptius exclusius de cada grup. Així, per exemple, els verbs amb l'infinitiu en *-er* i amb el radical accentuat poden pertànyer al model representat pel verb *témer*, que constitueix el grup 1, o al model representat pel verb *aparèixer*, que constitueix el grup 2. Per tant, també cal indicar a quin grup pertany cada lema, ja que no hi ha cap element a partir del qual es pugui preveure el tipus de flexió.

Segons la terminació i els models de conjugació que s'han pogut observar, s'han establert els grups verbals següents:

Segona conjugació

Grup 1: Verbs amb l'infinitiu acabat en *-er* i amb el radical accentuat (ex. *témer*)

Grup 2: Verbs amb l'infinitiu acabat en *-ixer* (ex. *aparèixer*)

Grup 3: Verbs amb l'infinitiu acabat en *-éixer* o *-àixer* (ex. *néixer* o *nàixer*)

Grup 4: Verbs amb l'infinitiu acabat en *-er* i sense accent en el radical (ex. *desfer*)

Grup 5: Altres verbs amb l'infinitiu acabat en *-er* (ex. *poder*)

Grup 6: Verbs amb l'infinitiu acabat només en *-re* (ex. *perdre*)

Grup 7: Verbs amb l'infinitiu acabat en *-re* o *-er* (ex. *cabre* o *caber*)

Grup 8: Verbs amb l'infinitiu acabat només en *-dre* (ex. *vendre*)

Grup 9: Verbs amb l'infinitiu acabat en *-dre* o *-er* (ex. *caldre* o *caler*)
Grup 10: Verbs amb el mateix paradigma que el grup anterior, però amb l'infinitiu acabat només en *-er* (ex. *prevaler*)
Grup 11: Verbs amb l'infinitiu acabat en *-ure* (ex. *plaure*)
Grup 12: Verbs amb l'infinitiu acabat en *-eure* o *-aure* (ex. *jaure* o *jaure*)
Grup 13: Altres verbs amb l'infinitiu acabat en *-re* (ex. *veure*)
Grup 14: Verbs amb l'infinitiu acabat en *-r* (ex. *dur*)

Tercera conjugació

Grup 1: Verbs amb l'infinitiu acabat en *-ir* i amb increment (ex. *patir*)
Grup 2: Verbs amb l'infinitiu acabat en *-ir* sense increment (ex. *ajupir*)
Grup 3: Verbs amb l'infinitiu acabat en *-ir* o *-dre* (ex. *tenir* o *tindre*)

Com es pot observar, la segona conjugació presenta un grau de complexitat molt superior a la tercera conjugació. En tots dos casos, però, cal indicar per a cada entrada lèxica o lema quin és el segment variable que cal substituir pels segments dels quadres de conjugació corresponents. I encara hi ha molts verbs que no s'adiuen completament a cap dels grups indicats. Vegem-ne un exemple:

Verbs amb l'infinitiu acabat en -er[5]
Part que s'ha de substituir: -ER

Grup 1: TÉMER, complànyer, esprémer, fènyer, fúmer, plànyer, terratrémer, trémer (llista exhaustiva)

inf.	[tém]er
ger.	ent
part.	ut, uda, uts, udes

5. Tots els verbs d'aquest grup tenen el radical només accentuat en l'infinitiu; en la resta dels casos, el radical és, si no s'indica el contrari, el mateix sense accent.

pr. ind.	o	s	—	em	eu	en
pret. imperf.	ia	ies	ia	íem	íeu	ien
pret. perf.	í	eres	é	érem	éreu	eren
fut.	eré	eràs	erà	erem	ereu	eran
cond.	eria	eries	eria	eríem	eríeu	erien
pr. subj.	i	is	i	em	eu	in
pret. subj.	és	essis	és	éssim	éssiu	essin
imp.		—	i	em	eu	in

PÉIXER (llista exhaustiva)
Diferències respecte al grup 1:

pr. ind.	o	**es**	—	em	eu	en

PERTÀNYER (llista exhaustiva)
Diferències respecte al grup 1:

part. **ut, uda, uts, udes** // [**pertang**] **ut, uda, uts, udes**

ATÈNYER, empènyer, espènyer (llista exhaustiva)
Diferències respecte al grup 1:

part. [**at**]**ès, esa, esos, eses**

CONSTRÈNYER, destrènyer, estrènyer, restrènyer (llista exhaustiva)
Diferències respecte al grup 1:

part. [**constr**]**et, eta, ets, etes**

VÈNCER, colltòrcer, contòrcer, convèncer, destòrcer, distòrcer, estòrcer, retòrcer, revèncer, tòrcer (llista exhaustiva)
Diferències respecte al grup 1:

part. [venç]ut, uda, uts, udes

pr. ind.	[venç]o	**es**	[venç]-	em	eu	en
imp.		[venç]-	i	em	eu	in

CRÉIXER, acréixer, decréixer, desmerèixer, irèixer, merèixer, recréixer, sobrecréixer (llista exhaustiva).

Diferències respecte al grup 1:

part. [**cresc**]ut, uda, uts, udes

pr. ind.	o	**es**	—	em	eu	en
pret. perf.	í / [**cresqu**]í	eres / [**cresqu**]eres	é / [**cresqu**]é	érem / [**cresqu**]érem	éreu / [**cresqu**]éreu	eren / [**cresqu**]eren
pr. subj.	i	is	i	em / [**cresqu**]em	eu / [**cresqu**]eu	in
pret.subj.	és / [**cresqu**]és	essis / [**cresqu**]essis	és / [**cresqu**]és	éssim / [**cresqu**]éssim	éssiu / [**cresqu**]éssiu	essin / [**cresqu**]essin
imp.		—	i	em / [**cresqu**]em	eu	in

Implementació

Les regularitats observades en els verbs de la primera conjugació enfront de la diversitat de paradigmes flexius no previsibles de la segona i la tercera, ha obligat a fer-ne tractaments diferents. La primera conjugació es resol amb 10 regles de dos nivells, com ara:

regla c_per_qu: {[ ] [q,u] [sufix-primera] ⇐ [ ] [c,+] [sufix-prime ra]}

Els verbs de la segona i tercera conjugacions, en canvi, no es tracten amb regles de dos nivells, ja que no hi ha regularitats suficients. Per contra, els verbs d'aquestes conjugacions s'han agrupat en 18 models de 52 formes cadascun. Si tenim en compte que la GP de la flexió verbal consta d'una sola regla de formació de paraules (que combina

una arrel amb un sufix), i multipliquem els 18 models per les 52 formes que contenen cadascun, obtindrem 936 sufixos diferents. La GP, però, és una gramàtica d'unificació que ha permès reagrupar sufixos de la mateixa forma que pertanyen a models diferents; d'aquesta manera el nombre de sufixos es redueix a 386.

Altres analitzadors morfològics per al català

En aquesta secció compararem breument CATMORF amb altres aproximacions computacionals a la morfologia catalana.

Hernàndez 1992

HERNÀNDEZ (1992) presenta un sistema per a l'ensenyament de la flexió verbal del català; la seva cobertura és, doncs, limitada. En aquest sistema s'assumeix que cada forma verbal està composta per una arrel i un sufix verbal; per tant, analitzar una forma verbal (i deduir-ne la informació morfològica associada) consisteix a trobar una combinació vàlida d'arrel amb una terminació verbal. Totes les combinacions possibles d'arrels i sufixos verbals estan agrupades en 94 models.

Un dels problemes principals d'aquest sistema és que no s'han agrupat (o subespecificat) els diferents sufixos que pertanyen a més d'un model; com a conseqüència, el sistema té un nombre molt elevat de terminacions verbals (unes 5.000).

El nostre tractament de la flexió verbal és molt similar pel que fa a la segona i tercera conjugació, però bastant diferent pel que fa a la primera. En aquest últim cas, la diferència es troba en el fet que a CATMORF hi ha un únic model per a la primera conjugació de 10 regles per a tractar els diferents canvis ortogràfics. D'altra banda, CATMORF agrupa terminacions que pertanyen a diferents paradigmes, de manera que tenim en total 386 sufixos verbals.

Martí 88

MARTÍ (1988) descriu tots els models necessaris per a tractar la flexió nominal, la flexió verbal i la derivació, però només s'incorporen mostres representatives de cada paradigma flexiu o derivatiu. El seu analitzador morfològic és un autòmat markovià ampliat amb condicions; els elements que el componen són els següents:

1) Diccionari d'arrels i sufixos (tant flexius com derivatius)
2) Un conjunt de regles que constitueixen l'autòmat i que validen la concatenació d'arrels amb sufixos. Les regles poden incorporar restriccions sobre la informació morfològica associada a les unitats dels diccionaris.
3) Un conjunt de models en que s'agrupen les arrels i els sufixos.
4) Els atributs morfològics associats a les unitats dels diccionaris i als models.

Analitzar una paraula consisteix a segmentar-la d'esquerra a dreta, trobar tots els segments resultants al diccionari i validar-ne la concatenació mitjançant l'autòmat.

Des del punt de vista teòric, la cobertura de MARTÍ (1988) és superior, perquè incorpora la derivació. Des del punt de vista pràctic, però, el lèxic de CATMORF és molt més ampli i inclou molts derivats lexicalitzats (per ex. *caseta*).

de Yzaguirre 95

YZAGUIRRE (1995) consisteix en una base de dades de formes flexionades generades a partir de les entrades que es troben al DEC83. També inclou algunes regles que tracten alguns casos de prefixació i els adverbis acabats en *-ment*. Tot i que l'anàlisi es fa a partir de les formes flexionades, la incorporació de noves formes es fa mitjançant l'assignació del lema a algun model de flexió. El grau de cobertura d'aquest sistema i CATMORF és pràcticament idèntic; la diferència principal es troba en els diferents models existents.

FITXA TÈCNICA

Les característiques tècniques més rellevants de CATMORF són:

1) El sistema cobreix la flexió nominal i verbal, a més d'alguns processos derivatius, alguns casos de prefixació i els adverbis acabats en *-ment.*
2) 114 regles donen compte de la flexió nominal i 10 regles donen compte de la flexió verbal de la primera conjugació.
3) La GP té 1 regla de formació de paraules per tractar la flexió verbal i 15 regles per tractar la flexió nominal.
4) Hi ha dos lèxics disponibles:
   DIEC (amb unes 70.500 entrades)
   DEC83 + DIEC + ampliacions DEC83 (amb unes 85.000 entrades)
5) El sistema s'ha implementat en Prolog i corre sobre plataformes UNIX.

TONI BADIA, ÀNGELS EGEA, TONI TUELLS

BIBLIOGRAFIA

J. ALLEN (1995): *Natural Language Understanding*, Califòrnia, The Benjamin/Cummings Publishing Company, Inc.

CATMORF, 1997: T. BADIA, A. EGEA i T. TUELLS (1997): *CATMORF: Multi two-level steps for Catalan morphology,* dins *Demo Proceedings of Applied Natural Language Processing (ANLP) 97*, Washington.

DEC83: ENCICLOPÈDIA CATALANA (1983): *Diccionari de la llengua catalana*.

DIEC: INSTITUT D'ESTUDIS CATALANS (1995): *Diccionari de la llengua catalana*.

G. GAZDAR, C. MELLISH (1989): *Natural language Processing in PROLOG. An Introduction to Computational Linguistics*, Wokingham, Addison-Wesley Publishing Company.

L. HERNÀNDEZ (1992): *Un sistema per a l'anàlisi automàtica de la flexió del verb català*, Tesi doctoral, Universitat Politècnica de Catalunya.

R. KAPLAN i M. KAY (1994): *Regular Models of phonological rule systems,* dins «Computational Linguistics», núm. 20 (3): ps. 331-378.

K. KOSKENNIEMI (1984): *A General Computational Model for Word-form Recognition and Production,* dins *Proceedings of the 10th International Conference on Computational Linguistics (COLING-84*), Stanford.

M. A. MARTÍ (1988): *Processament del Llenguatge Natural: un sistema d'anàlisi morfològica per ordinador*, Tesi doctoral, Departament de Filologia Romànica, Facultat de Filologia, Universitat de Barcelona.

A. GAL et al. (1991): *PROLOG for Natural Language Processing*, John Wiley & Sons. UK.

SEGMORF, 1996: T. BADIA , A. EGEA i T. TUELLS (1996): *SEGMORF: Un formalismo para analizadores morfológicos de dos niveles*, dins *Actas del XII Congreso de la Sociedad Española Para el Procesamiento del Lenguaje Natural (SEPLN*), Sevilla.

SEGMORF, 1997: T. BADIA i T. TUELLS (1997): *SEGMORF: An extension of the Alep morphographemic segmentation formalism*, dins *Proceedings of the 3rd Alep User Group Workshop*, Saarbrücken.

R. SPROAT (1992): *Morphology and Computation*, MIT Press.

T. TUELLS (1998): *Constructing and Updating the Lexicon of a Two-level Morphological Analyzer from a Machine-Readable Dictionary*, dins *Proceedings of the First Language Resources and Evaluation (LREC-98) Conference*, Granada.

L. DE YZAGUIRRE (1995): Memòria presentada al concurs de mèrits de la Universitat Pompeu Fabra (ref: UPF 153TM).