Introducció a la lingüística computacional

Xavier Gómez Guinovart (Universidade de Vigo)
Antoni Oliver Gonzàlez (Universitat Oberta de Catalunya)

PID_00155233

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Introducció

En aquest mòdul presentem una introducció a la lingüística computacional, amb l'objectiu de definir aquesta àrea d'estudi i presentar les línies de recerca més rellevants. En un primer apartat es pretén definir l'àmbit d'aquesta àrea d'estudi i les seves principals línies d'investigació i desenvolupament: la lingüística computacional teòrica, la lingüística computacional aplicada i la informàtica aplicada a la traducció. Després es presenten els principals models i formalismes lingüístics computacionals amb exemples concrets que el lector podrà provar. El tercer apartat està dedicat a presentar les principals aplicacions de la lingüística computacional: comprensió i generació del llenguatge natural, tecnologies de la parla, processament documental, traducció automàtica. A continuació es presenten les principals aplicacions de la informàtica aplicada a la lingüística: la lingüística de corpus i la lingüística històrica computacional.

Aquest mòdul és una traducció, adaptació i actualització de l'obra de Gómez Guinovart (2000). Tot i que s'ha seguit l'estructura bàsica del document original, s'ha intentat actualitzar al màxim les referències i adaptar els exemples a la llengua catalana.

Objectius

En els materials didàctics d'aquest mòdul presentem els continguts i les eines imprescindibles per a assolir els objectius següents:

1. Definir l'àrea d'estudi de la lingüística computacional.
2. Comprendre el caràcter multidisciplinari d'aquesta àrea d'estudi.
3. Presentar les línies de recerca més rellevants d'aquesta àrea d'estudi.
4. Presentar de manera clara els principals models i formalismes lingüístics computacionals, amb exemples clars i que el lector pot reproduir, provar i modificar.

1. Àmbit de la lingüística computacional

1.1. Línies d'investigació

La *lingüística computacional* és un àmbit científic interdisciplinari vinculat a la lingüística i a la informàtica, i que té com a objectiu incorporar als ordinadors l'habilitat de tractar el llenguatge natural humà i facilitar el tractament informàtic i l'estudi de les llengües.

Tot i ser una disciplina relativament recent, per poder delimitar-ne el camp d'estudi cal diferenciar un mínim de tres línies d'investigació i desenvolupament principals:

- la lingüística computacional teòrica,
- la lingüística computacional aplicada i
- la informàtica aplicada a la lingüística.

Dins del primer vessant, la *lingüística computacional teòrica*, es poden distingir com a mínim tres objectius complementaris:

- l'elaboració de models lingüístics formals que siguin implementables computacionalment,
- l'aplicació d'aquests models a algun dels nivells de descripció de la lingüística i
- la comprovació automatitzada de l'adequació d'una teoria lingüística i de les seves prediccions.

La confecció de models computacionals del llenguatge mitjançant formalismes lingüístics adequats i la utilització d'aquests models en la descripció lingüística faciliten la detecció d'errors i incoherències en la descripció dels fenòmens lingüístics i ofereixen un mitjà pràctic i efectiu per a observar la interacció de tots els components del model.

En segon lloc, l'orientació més tecnològica de la lingüística computacional, la *lingüística computacional aplicada*, s'orienta al disseny i elaboració de sistemes informàtics que siguin capaços de comprendre, produir i traduir enunci-

ats orals i escrits en llenguatge natural. Aquesta orientació es concreta en el desenvolupament d'aplicacions que es poden classificar en quatre categories:

- els sistemes de comprensió i generació d'enunciats (com els programes de consulta en llenguatge natural a bases de dades i els sistemes automàtics de diàleg per línia telefònica);

- les aplicacions de les tecnologies de la parla (com els programes de dictat i els sistemes de conversió de text a veu);

- les eines de processament documental per a l'elaboració, gestió i revisió de documents (com els programes de verificació de la correcció lingüística de textos, els programes de generació automàtica de resums, els sistemes d'extracció d'informació, els sistemes de recuperació d'informació i els programes de catalogació documental automatitzada);

- les eines de processament plurilingüe, en el seu doble vessant d'aplicacions didàctiques per a l'ensenyament de llengües (com els mètodes d'aprenentatge d'idiomes assistit per ordinador i els programes de creació d'exercicis de llengua) i d'eines d'ajut a la traducció (com els programes de traducció automàtica, les bases de dades terminològiques i els programes de memòries de traducció).

Aquest camp rep diverses denominacions, depenent de l'activitat que es vulgui destacar: processament del llenguatge natural, tecnologies de la llengua o enginyeria lingüística.

Denominacions de la lingüística computacional aplicada

La lingüística computacional aplicada rep diverses denominacions: processament del llenguatge natural, tecnologies de la llengua o enginyeria lingüística.

Per últim, el camp de treball caracteritzat per l'aplicació dels ordinadors a la investigació lingüística, és a dir, a l'estudi científic del llenguatge i de les llengües, acostuma a rebre el nom d'*informàtica aplicada a la lingüística* o *lingüística informàtica*. El terme es pot aplicar en sentit ampli a totes les subdisciplines de la lingüística que fan servir eines informàtiques, tot i que en general es reserva aquest terme a aquelles àrees d'investigació on aquestes eines informàtiques tenen una incidència més gran, com poden ser la *lingüística de corpus* o la *lingüística històrica computacional*.

1.2. Dimensions interdisciplinària i social

Des del punt de vista de la seva vinculació amb la informàtica, i també per motius històrics, la lingüística computacional està considerada una subdisciplina de la *intel·ligència artificial*, una especialitat de la informàtica que s'ocupa de la comprensió de la intel·ligència i del disseny de màquines intel·ligents, és a dir, màquines i programes que presenten característiques associades amb l'enteniment humà, com la raó, la comprensió del llenguatge parlat i escrit, l'aprenentatge o la presa de decisions.

Tanmateix, des del punt de vista de la seva relació amb la lingüística, la lingüística computacional també es pot considerar un subdisciplina de la *lingüística teòrica*, ja que un dels seus objectius és l'elaboració de models formals del llenguatge humà que es puguin implementar computacionalment. En aquest sentit, la lingüística computacional està estretament relacionada amb la *psicolingüística* i amb la *lingüística cognitiva*, pel seu interès compartit en la descripció i modelatge de l'activitat mental implicada en el processament lingüístic.

Finalment, com a disciplina lingüística experimental, la lingüística computacional constitueix l'àrea de treball de la *lingüística aplicada* específicament interessada en aplicar els resultats i mètodes de la la investigació lingüística a l'elaboració de productes comercials i d'investigació en el marc de les indústries de la llengua. L'ampli ventall d'aplicacions lingüístiques de la informàtica enllaça la lingüística computacional amb les diferents disciplines lingüístiques i no lingüístiques relacionades amb cada una de les aplicacions, com l'*enginyeria de telecomunicacions* (en relació amb les aplicacions de les tecnologies de la parla), les *ciències de la documentació* (amb els sistemes de gestió documental), la *traductologia* (amb les eines d'ajut a la traducció), la *didàctica de les llengües* (amb l'ensenyament de llengües assistit per ordinador), l'*anàlisi del discurs* (amb els sistemes de diàleg) o la *lexicografia* (amb els diccionaris electrònics).

Juntament amb aquesta dimensió interdisciplinària de la disciplina, cal destacar també la seva dimensió social en una societat de la informació cada cop més global i més influenciada per les telecomunicacions i els seus interessos culturals i comercials. Les aplicacions actuals de la lingüística informàtica en les telecomunicacions posen de manifest una clara tendència a permetre l'ús de la llengua pròpia per a accedir a totes les possibilitats d'informació, comunicació i consum que estan a l'abast de les persones que habiten el denominat "primer món" per les noves tecnologies. Per la seva gran incidència social, la presència d'una determinada llengua en aquest àmbit és determinant per assolir o conservar l'estat de llengua normalitzada (Comissió Europea 1998, 14-15).

2. Models i formalismes lingüístics

Un dels interessos centrals de la investigació en lingüística computacional és la implementació informàtica de teories lingüístiques i l'elaboració de models computacionals del llenguatge.

A continuació presentem algunes de les orientacions més destacables d'aquestes dues línies de treball complementàries, en què ens limitem als nivells lèxic i sintàctic, als models simbòlics i als formalismes d'unificació. Altres línies de treball actuals destacades són la *fonologia computacional* (Bird, 1995), les *xarxes lèxicosemàntiques* (Wanner, 1996; Alonge i altres, 1998), la *semàntica computacional* (Rosner i Johnson, 1996) i els *models lingüístics probabilístics* (Charniak, 1993).

2.1. Models lingüístics computacionals

Un dels objectius fonamentals de la lingüística computacional és el desenvolupament de teories lingüístiques formals i implementables informàticament. Aquestes teories constitueixen models computacionals del funcionament del llenguatge i reben el nom de *models lingüístics* (Shieber, 1998). Dins d'aquesta línia d'investigació cal destacar la *gramàtica funcional lèxica* o LFG (*Lexical-Functional Grammar*) (Bresnan, 1999), la *gramàtica d'estructura sintagmàtica ampliada* o GPSG (*Generalised Phrase Structure Grammar*) (Gazdar i altres, 1985), la *gramàtica d'estructura sintagmàtica regida pel nucli* o HPSG (*Head-Driven Phrase Structure Grammar*) (Pollard i Sag, 1994) i la *gramàtica categorial* (Solias, 1996), models agrupats genèricament sota la denominació de *gramàtiques d'unificació* (Shieber, 1986; Ruiz Antón, 1996; Balari Ravera, 1999) ja que recorren a aquest procediment matemàtic en les seves descripcions lingüístiques.

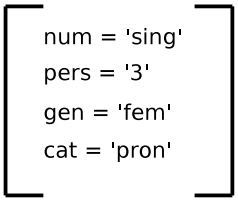
En la majoria d'aquests models els objectes lingüístics estan representats en forma d'objectes matemàtics denominats *estructures de trets* (ET) i els fenòmens lingüístics es descriuen formulant equacions amb aquestes ET. La unificació és l'operació matemàtica que permet resoldre aquests sistemes d'equacions, combinant les informacions lingüístiques codificades en les ET que hi ha associades. Cada ET està formada per una matriu de trets, i cada tret consisteix en una parella [Atribut=Valor] que especifica un paràmetre lingüístic i el valor que adopta aquest paràmetre, com per exemple [número=singular] o

Unificació

La unificació és un mecanisme o operació matemàtica que permet compondre constituents, és a dir, combinar informació sempre que aquesta sigui compatible.

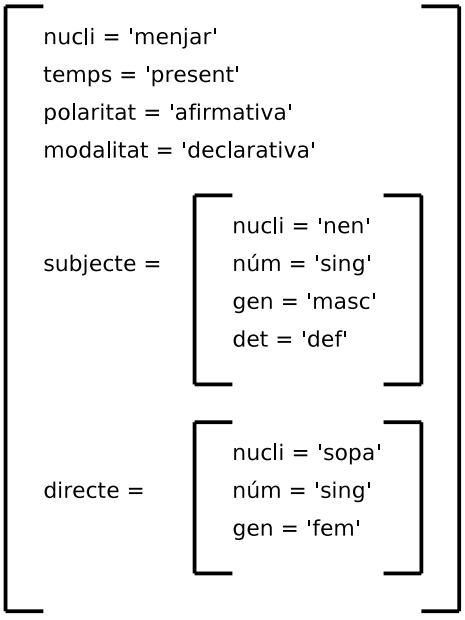
[persona=tercera]. Així, una manera de representar la informació lingüística associada amb el pronom "ella" seria mitjançant l'ET de la figura 1.

Figura 1. Estructura de trets per al pronom "ella"



El valor d'un tret pot ser un altre tret, de manera que es converteix en un tret complex. Per exemple, l'oració "el nen menja sopa" es pot representar per l'ET de la figura 1, on els trets per al subjecte i el complement directe són trets complexos (figura 2).

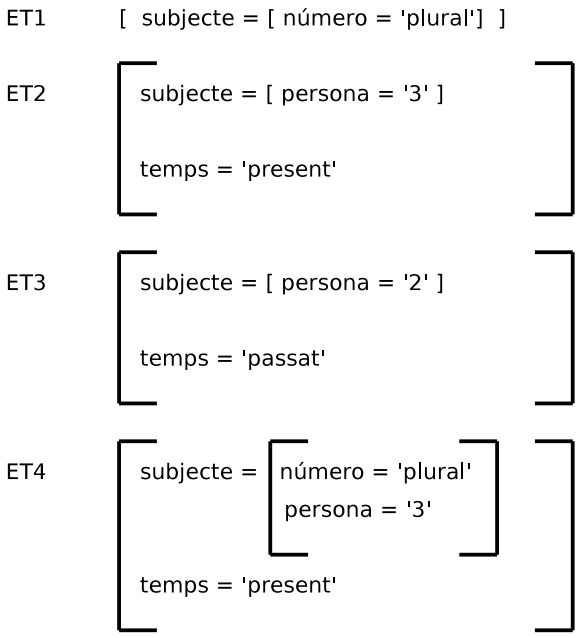
Figura 2. Estructura de trets oracional amb trets complexos



La unificació és una operació que combina dues ET i produeix com a resultat una altra ET que té tota la informació continguda en les dues ET originals, sempre que la informació no sigui contradictòria.

Si la informació continguda en una de les ET resulta contradictòria amb la continguda en l'altra, no es podrà portar a terme la unificació. Per exemple, en la figura 3, l'ET1 i l'ET2 unifiquen en l'ET4, però l'ET3 no pot unificar-se ni amb l'ET1 ni amb l'ET2.

Figura 3. Exemple d'unificació d'estructures de trets

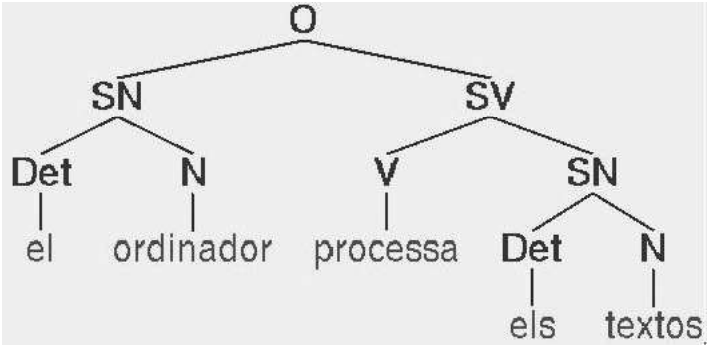


Per una altra banda, en la formalització de models lingüístics el procediment més emprat per a la descripció de les estructures de constituents admeses en una llengua són les *regles sintagmàtiques* (també denominades *regles d'estructura de constituents* o *regles de reescriptura*). Aquestes regles adopten la forma $A \rightarrow B$, on A representa una categoria sintàctica i B són els constituents immediats de A. Per exemple, una regla que expressaria una estructura sintagmàtica de moltes oracions del català podria ser $O \rightarrow SNSV$ ("una oració està formada per un sintagma nominal seguit d'un sintagma verbal"). Bàsicament, una gramàtica d'estructura sintagmàtica és un conjunt de regles sintagmàtiques que descriuen les estructures sintàctiques acceptables d'una llengua. L'exemple de la figura 4 podria constituir un fragment d'una gramàtica sintacmàtica del català que descriuria, entre altres, l'estructura de constituents representada mitjançant el diagrama arbori de la figura 5. Fixeu-vos que en aquesta gramàtica no hem tractat l'apostrofació de l'article. L'apostrofació es podria tractar afegint una entrada lèxica corresponent a l'article apostrofat, o bé fent un preprocessat i postprocessat que tractés l'apostrofació.

Figura 4. Fragment de la gramàtica d'estructura sintagmàtica

```
O –> SN SV

SN –> Det N
SV –> SN

Det –> "el"
Det –> "els"
N –> "ordinador"
N –> "ordinadors"
N –> "text"
N –> "textos"
V –> "processa"
V –> "processen"
```

Figura 5. Estructura de constituents descrita per la gramàtica



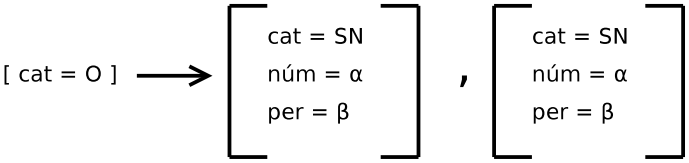
Les regles sintagmàtiques augmentades permeten caracteritzar els constituents i establir condicions d'igualtat entre les seves propietats. Per exemple, es pot augmentar amb condicions el fragment de la gramàtica sintagmàtica presentada anteriorment per a incorporar l'obligatorietat de la concordança de número i persona entre el verb i el seu subjecte, modificant la primera regla com es mostra a la figura 6.

Figura 6. Regla sintagmàtica augmentada

```
O –> SN SV
<SN_num = SV_num>
<SN_per = SV_per>
```

Aquesta regla ens indica que una oració està formada per un SN seguit d'un SV, i que el número i la persona del SN i del SV coincideixen. Substituint els símbols categorials indivisibles per estructures de trets, com a la figura 7, es poden redefinir aquestes equacions en forma d'igualtat de variables.

Figura 7. Estructures de trets i regles sintagmàtiques augmentades



Finalment, assignant al SN la funció de subjecte i al SV la de predicat, i fent servir aquestes dues funcions com a atributs de trets complexos, es podria reduir la regla a una simple estructura de trets (figura 8).

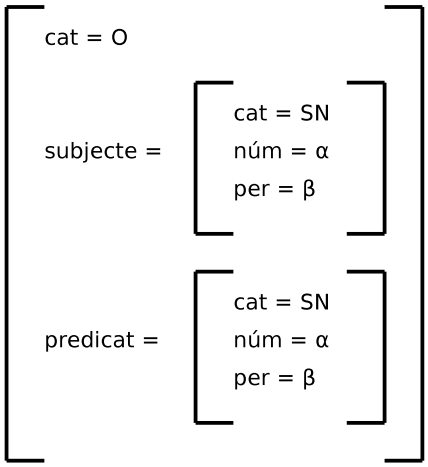
Aquesta perspectiva estàtica de la manipulació dels objectes lingüístics permet simplificar el component sintàctic i desplaçar cap al lèxic la informació lingüística que ha estat tradicionalment tractada en la gramàtica, cosa que implica una més gran complexitat de l'organització i el contingut del component lèxic. Com a resultat d'aquest desplaçament, en els models d'orientació lexicalista les regles sintagmàtiques són inexistents (com per exemple en la *gramàtica categorial*) o de natura molt general (com en l'HPSG), cosa que habitualment implica adoptar alguna de les versions de la *teoria X amb barra*, on el nucli del constituent, caracteritzat en el lèxic, proporciona pràcticament tota

Teoria de la X amb barra

La teoria de la X amb barra garanteix que els elements lèxics es projecten adequadament en l'estructura sintagmàtica, de manera que es defineixen els sintagmes com a projeccions dels nuclis lèxics.

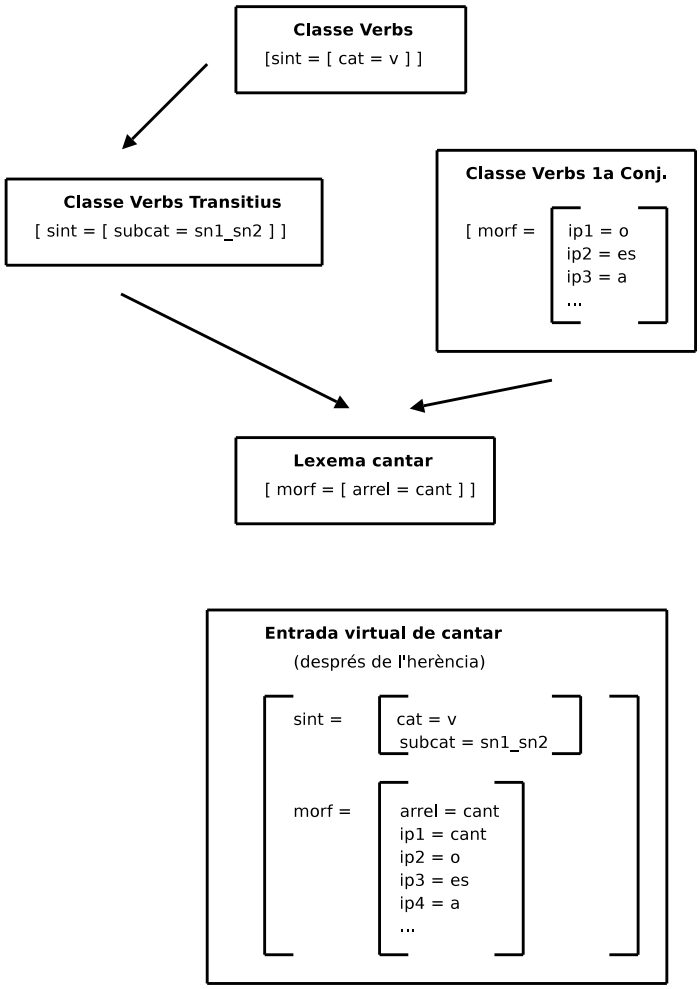
la informació sobre les propietats sintàctiques i semàntiques dels seus complements. Una simplificació addicional de la gramàtica consisteix a distingir entre les regles que representen l'ordre seqüencial dels constituents (*regles de precedència lineal*) i les que representen les seves relacions de dependència estructural o jeràrquica (*regles de domini immediat*), distinció practicada i difosa per la GPSG i l'HPSG.

Figura 8. Estructura de trets amb variables



El component lèxic d'aquests models s'organitza com una xarxa de nodes jerarquitzada amb herència múltiple, on cada node està format per una estructura de trets corresponent a un lexema o a una classe de lexemes. Els nodes per a les classes de lexemes contenen totes les propietats lingüístiques que són comunes a la classe. Per exemple, la classe "verbs regulars de la primera conjugació" del lèxic català podria contenir la informació morfològica necessària per a flexionar els verbs d'aquesta classe; la classe "verbs transitius", l'especificació del context sintàctic característic dels verbs d'aquesta categoria, i la classe "verbs", l'adscripció categorial compartida per tots els verbs de la llengua. Els nodes de les xarxes lèxiques poden heretar una part de les seves propietats dels nodes de la jerarquia de què depenen. Per exemple, el node de la classe "verbs transitius" pot heretar la seva categoria sintàctica del node de la classe "verbs", i el node del lexema "cantar" pot heretat les seves propietats categorials i de subcategorització del node de la classe "verbs transitius". A més a més, l'herència pot venir de diferents nodes, sempre que les propietats heretades no siguin contradictòries. D'aquesta manera, el node del lexema "cantar" podria heretar les seves característiques sintàctiques del node de la classe "verbs transitius" i les característiques morfològiques de la classe de "verbs regulars de la primera conjugació" (figura 9). D'aquesta manera, s'aconsegueix eliminar del lèxic la informació lingüística redundant, ja que no és necessari repetir la descripció de la subcategorització transitiva en totes les entrades dels verbs transitius, ni el comportament flexiu de la primera conjugació en totes les entrades dels verbs d'aquest paradigma morfològic.

Figura 9. Representació de la informació lèxica



2.2. Formalismes lingüístics

Els *formalismes lingüístics* (o *sistemes de programació lingüística*) són llenguatges artificials dissenyats per a representar la informació lingüística. Alguns formalismes lingüístics, com per exemple DCG (*Definite Clause Grammar*) (Pereira i Warren, 1980), FUG (*Functional Unification Grammar*) (Kay, 1982), PATR (Shieber, 1986), DATR (Evans i Gazdar, 1996), la morfologia de dos nivells (Koskenniemi, 1983) o ALE (Carpenter i Penn, 1997) poden ser interpretats pels ordinadors, per la qual cosa són especialment adequats per a la implementació informàtica i la verificació automàtica de les teories lingüístiques. Sovint es fan servir també llenguatges de programació de propòsit general, com el Prolog (Gazdar i Mellish, 1989). A continuació presentarem dues aplicacions simples, que il·lustren els mètodes de programació lingüística en els nivells d'anàlisi sintàctica i morfològica, mitjançant els formalismes PATR i DATR.

PATR és un formalisme dissenyat per a escriure gramàtiques d'estructura sintagmàtica augmentada amb estructures de trets sobre les quals opera la unificació. PC-PATR* és un programa que permet implementar informàticament gramàtiques escrites en aquest formalisme. Per convertir en PC-PATR el frag-

***ftp://ftp.sil.org/software/dos/pcpatr138.zip**

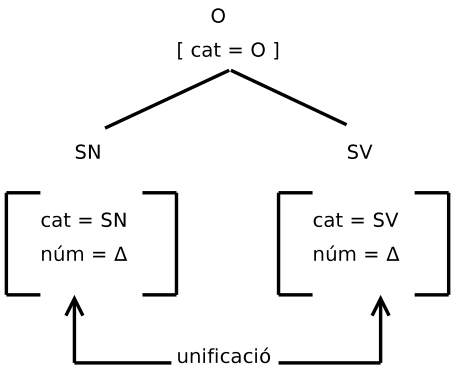
ment anterior de gramàtica d'estructura sintagmàtica del català (afegint concordança nominal i verbal), primer cal crear un fitxer que contingui la gramàtica, com per exemple el fitxer patr01.grm (figura 10).

Figura 10. Gramàtica patr01.grm en PC-PATR

```
Rule O -> SN SV
<SN num> = <SV num>
Rule SV -> V SN
<SV num> = <SN num>
Rule SN -> Det N
<SN num> = <N num>
<Det gen> = <N gen>
<Det num> = <N num>
```

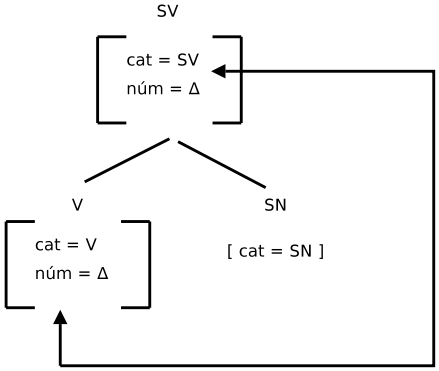
En PC-PATR, les regles han d'anar precedides per la paraula anglesa RULE. La primera regla de l'exemple defineix una estructura de constituents augmentada amb trets, en les quals opera la unificació sobre els trets de número nominal i verbal (figura 11). Si aquests trets fossin contradictoris, no es podria aplicar la unificació i, per tant, no es podria portar a terme l'anàlisi de la seqüència.

Figura 11. Unificació del número nominal i verbal



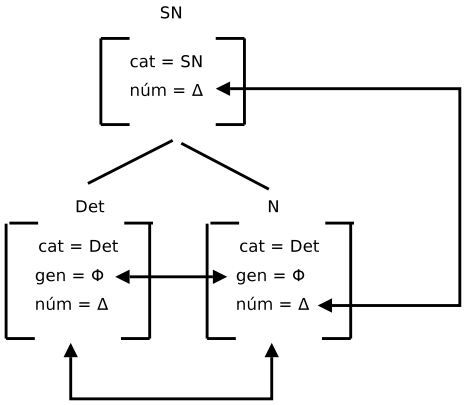
En l'estructura sintàctica definida per la segona regla, la unificació s'aplica als trets de número verbal, de manera que el número del verb sigui també el número del seu sintagma verbal (figura 12). Mitjançant aquesta elevació del valor de número de V a SV, s'obté un SV amb el número d'acord amb el seu nucli, cosa que permet verificar la concordança amb el subjecte expressada en la primera regla.

Figura 12. Elevació de número per unificació



Finalment, en la tercera regla, es construeix l'estructura de constituents del SN, s'eleva el valor de gènere i número entre el determinant i el nom (figura 13).

Figura 13. Concordança i elevació de número



Un cop elaborada la gramàtica i emmagatzemada en el fitxer patr01.grm, cal definir les regles que introdueixen les peces lèxiques terminals en un fitxer, per exemple el patr01.lex (figura 14).

Figura 14. Lèxic patr01.lex en PC-PATR

```
\w el \\
\c Det \\
\f <gen> = m \\
   <num> = s \\ \\
\w els \\
\c Det \\
\f <gen> = m \\
   <num> = p \\ \\
\w text \\
\c N \\
\f <gen> = m \\
   <num> = s \\ \\
\w textos \\
\c N \\
\f <gen> = m \\
   <num> = p\\ \\
\w ordinador \\
\c N \\
\f <gen> = m \\
   <num> = s \\ \\
\w ordinadors \\
\c N \\
\f <gen> = m \\
   <num> = p\\ \\
\w processa \\
\c V \\
\f <num> = s \\ \\
\w processen \\
\c V \\
\f <num> = p\\
```

A partir d'aquestes definicions es creen un seguit d'estructures de trets per a inserir-les en una estructura sintagmàtica. Aquestes ET lèxiques contindran la informació declarada a PC-PATR amb el codi \w (de "word", paraula) convertida en un valor de lex (per lexema), la declarada amb el codi \c convertida en un valor de cat, i tota la resta de trets especificats pel codi \f (de "features", trets), com es mostra en l'ET de la figura 15 per a la paraula "ordinadors".

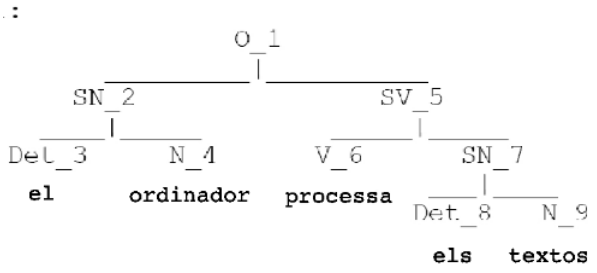
Figura 15. Concordança i elevació de número

lex	ordinadors
cat	N
gen	m
num	p

Donada la gramàtica de patr01.grm i el lèxic de patr01.lex, una interacció típica amb el sistema pot ser la representada en la figura 16, on PC-PATR porta a terme l'anàlisi sintàctica automàtica de la seqüència "l'ordinador processa els textos" i, com a resultat, construeix una estructura de constituents i indica les ET associades a cada node (numerats per tal que sigui més clar) de l'arbre sintàctic.

Figura 16. Anàlisi sintàctica automàtica amb PC-PATR

```
PC-PATR>load grammar patr01
Loading grammar from patr01.grm
PC-PATR>load lexicon patr01
Loading lexicon from patr01.lex
8 lexicon entries loaded from patr01.lex
PC-PATR>parse sentence: el ordinador processa els textos
```




```
O_1:
[ cat: O ]
SN_2: [ cat: SN
num: s ]
Det_3:
[ cat: Det
gen: m
lex: el
num: s ]
N_4:
[ cat: N
gen: m
lex: ordinador
num: s ]
SV_5:
[ cat: SV
num: s ]
V_6:
[ cat: V
lex: processa
num: s ]
SN_7:
[ cat: SN
num: p ]
Det_8:
[ cat: Det
gen: m
lex: els
num: p ]
N_9:
[ cat: N
gen: m
lex: textos
num: p ]
1 parse found
```

Per una altra banda, el llenguatge formal DATR permet implementar informàticament xarxes d'estructures de trets lèxics (de lexemes i classes de lexemes)

jerarquitzades i amb herència de propietats múltiples (és a dir, herència que pot venir de diferents nodes). La figura 17 recull l'exemple d'una xarxa lèxica del català il·lustrada a la figura 9, convertida en DATR i emmagatzemada en un fitxer (que hem anomenat herencia.dtr) per al seu processament posterior.

Figura 17. Fitxer de lèxic herencia.dtr en DATR

```
Verbs:
<sint cat> == v.
Verbs_Transitius:
<> == Verbs
<sint subcat> == sn1_sn2.
Verbs_1Conj: <morf ip1> == o
<morf ip2> == es
<morf ip3> == a.
cantar:
<morf arrel> == cant
<sint> == Verbs_Transitius
<morf> == Verbos_1Conj.
#show
<sint cat> <sint subcat> <morf arrel> <morf ip1> <morf ip2> <morf ip3>.
#hide Verbs Verbos_Transitius Verbos_1Conj.
```

D'acord amb les convencions de DATR, en herencia.dtr el noms dels nodes van seguits de dos punts (:), els trets de la seva ET apareixen entre els dos punts i el punt final (.), els trets (o cadenes de trets) de les estructures de trets van entre claudàtors triangulars (<>) i els seus valors s'indiquen amb dos signes igual seguits (==). La presència del nom d'un node com a valor d'un atribut indica l'origen de l'herència; per exemple, el tret <morf>==Verbs_1Conx del node "cantar" indica que aquest node hereta del primer les propietats de l'atribut <morf>. Els atributs buits (<>) es fan servir per a establir l'herència de tots els trets inclosos en el node especificat com a valor; la línia <>==Verbs del node Verbs_Transitius significa que aquest node hereta tots els trets de l'ET del node "Verbs" (en aquest cas, <sint cat>==v). Finalment, amb el codi #show apareixen els noms dels atributs que es volen visualitzar com a resultat del tractament del fitxer, i amb el codi #hide, els que es volen ocultar, però en referència amb els noms que hem fet sevir. Amb aquesta definició del lèxic i aquesta configuració, i fent servir el programa Q-DATR* per a processar-lo, l'ordinador pot realitzar l'avaluació automàtica de l'herència de propietats especificades en el lèxic i presentar els trets dels nodes de la xarxa que hem sol·licitat visualitzar (figura 18).

***ftp://ftp.cogs.sussex.ac.uk/pub/nlp/DATR/qdatr200.exe**

Figura 18. Resolució de l'herència amb Q-DATR

```
>> cp(c:\datr\herencia.dtr)
compiling c:\datr\herencia.dtr
6 sentences compiled >>
datr_theorem
cantar:
      <sint cat> = v
      <sint subcat> = sn1_sn2
      <morf raiz> = cant
      <morf ip1> = o
      <morf ip2> = es
      <morf ip3> = a.
```

3. Aplicacions de la lingüística computacional

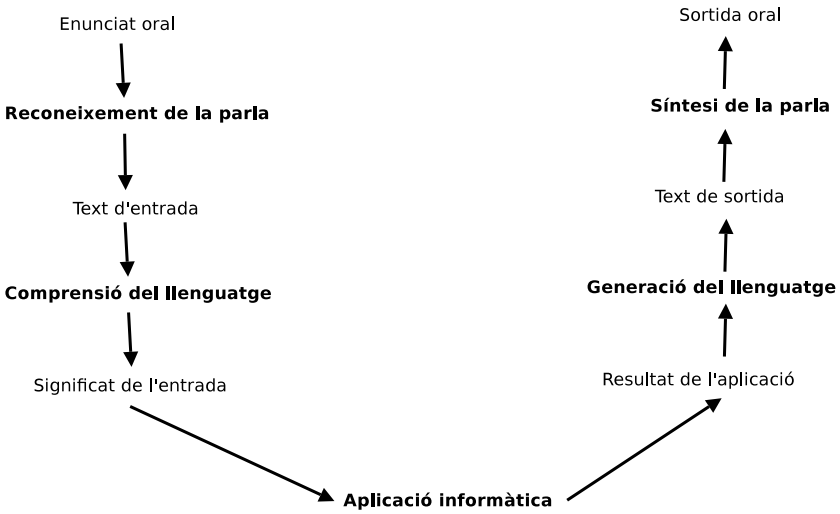
3.1. Comprensió i generació de llenguatge natural

Un dels objectius centrals de la lingüística computacional aplicada és permetre l'ús oral de la llengua materna com a mitjà de comunicació entre els ordinadors i les persones, amb la finalitat que les persones puguin accedir a totes les funcions dels ordinadors mitjançant ordres vocals expressades espontàniament amb el vocabulari i la sintaxi de la seva pròpia llengua i, al mateix temps, que els ordinadors presentin els resultats de les seves aplicacions en aquesta mateixa llengua de manera natural i comprensible per a les persones.

Els sistemes de comprensió del llenguatge natural són els programes informàtics que s'encarreguen de deduir el significat dels enunciats lingüístics d'entrada que processen, mentre que els sistemes de generació del llenguatge natural són els responsables de presentar els resultats de les aplicacions informàtiques en forma d'enunciats lingüístics (Allen, 1995; Reiter i Dale, 1997).

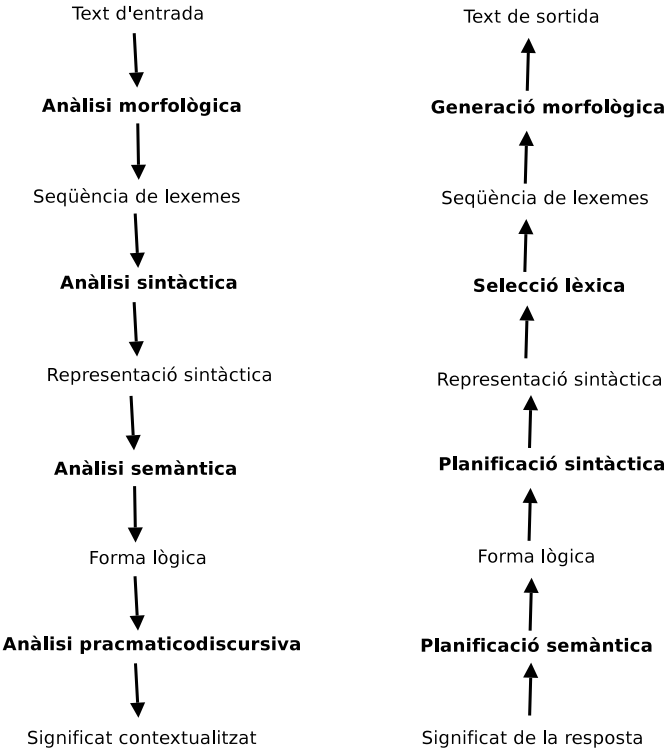
La combinació de les tècniques de comprensió i generació permet establir una interacció lingüística entre la persona i l'ordinador en situacions comunicatives ben delimitades, com les que es donen en els programes de consulta en llenguatge natural a bases de dades o en els sistemes automàtics de diàleg per línia telefònica (figura 19).

Figura 19. Interacció lingüística oral persona-ordinador



La generació i comprensió del llenguatge natural són tasques complexes que impliquen l'aplicació conjunta de tècniques molt diverses d'anàlisi i producció lingüística automàtica. Deixant de banda el processament del nivell fònic, que examinarem en l'apartat següent (dedicat específicament al reconeixement i la síntesi de la parla), el processament de la comprensió lingüística es realitza habitualment en quatre etapes successives que corresponen a l'anàlisi morfològica (etiquetador de categories i lematització), l'anàlisi sintàctica, l'anàlisi semàntica oracional i l'anàlisi pragmàtica i discursiva; mentre que la tasca de generació del llenguatge recorreria el camí invers, seguint les etapes de planificació semàntica, planificació sintàctica, selecció lèxica i generació morfològica (figura 20).

Figura 20. Components de la comprensió i generació del llenguatge



La utilitat operativa de les aplicacions de comprensió i generació del llenguatge està actualment circumscrita a àmbits d'interacció molt ben delimitats des d'un punt de vista temàtic i lingüístic, com poden ser la consulta telefònica a bases de dades dels horaris dels vols d'una companyia aèria mitjançant diàlegs dirigits per l'ordinador. Cal continuar treballant perquè el processament de la comprensió del llenguatge aconsegueixi nivells elevats de cobertura i precisió capaços de tractar de manera adequada els enunciats que apareixen en les interaccions lingüístiques espontànies. Un grau de cobertura alt implica que el programa de comprensió no deixi gairebé cap enunciat sense analitzar, mentre que un grau de precisió alt comporta que la majora dels enunciats reben l'anàlisi correcta. Una cobertura apropiada evitaria que els usuaris del sistema haguessin de repetir de diferents maneres un enunciat per indicació del programa, mentre que una bona precisió evitaria els errors d'interpretació per part del sistema.

En el camp de la generació del llenguatge, la investigació està centrada principalment en la generació del llenguatge escrit i en els entorns comunicatius d'interactivitat baixa. Per poder disposar en un futur de sistemes amb interacció lingüística oral espontània persona-ordinador, serà precís orientar els esforços d'investigació cap a la generació de llenguatge oral en diàlegs interactius, en què el programa de generació tingui en compte el contingut dels enunciats previs i adapti la seva producció lingüística a les intervencions de la persona interlocutora.

3.2. Tecnologies de la parla

Les *tecnologies de la parla* s'ocupen del processament dels aspectes fònics del llenguatge, amb l'objectiu de permetre la comunicació oral entre les persones i els ordinadors.

Segons la direcció de comunicació considerada, el tractament informàtic de la parla afronta dues tasques ben diferenciades: el processament de la percepció acústica o *reconeixement de la parla*, i el processament de la producció fonètica o *síntesi de la parla* (Dutoit, 1997; Llamas i Cardeñoso, 1997).

El reconeixement de la parla consisteix a convertir un enunciat oral en una cadena de símbols, com per exemple, un text escrit. La popularització de les tecnologies de reconeixement es deu als sistemes de dictat per a processament de text en ordinadors personals. Aquests programes de dictat, comercialitzats per empreses com IBM i Dragons Systems, ofereixen versions per a parla fragmentada, en què s'ha de fer una pausa entre les paraules, i versions per a parla contínua, que permeten dictar text sense necessitat de fer pauses entre les paraules.

Una de les característiques més desitjables en un sistema de reconeixement és la resistència al soroll ambient, amb la finalitat de poder-lo fer servir en entorns sorollosos (com en un fàbrica, per a controlar vocalment el braç d'un robot) o a través del telèfon (com per exemple per a dictar el número de telèfon a una central telefònica automàtica). Ara per ara, tot i l'interès evident que suscita aquesta qüestió entre els proveïdors de serveis de telecomunicacions, encara no hi ha una solució definitiva a causa de la baixa fiabilitat del reconeixement en entorns sorollosos (Tapias Merino, 1999).

Una altra de les dificultats del reconeixement de la parla contínua sense restriccions consisteix a reconèixer la parla amb independència de la persona. Parlem de reconeixement sense restriccions quan el sistema reconeix el vocabulari general d'una llengua. Això és imprescindible en els sistemes de dictat,

tot i que en altres aplicacions, com les centrals telefòniques automatitzades, es poden limitar a reconèixer unes poques paraules. Tanmateix, una central automatitzada d'un sistema públic de consulta telefònica ha de ser capaç de reconèixer la parla de qualsevol persona que truqui, mentre que un sistema de dictat es pot especialitzar en les característiques de la parla d'una persona concreta. Donat que el reconeixement de parla contínua sense restriccions amb independència de la persona no ha assolit encara un grau de fiabilitat acceptable per a la seva comercialització, els sistemes de dictat per a parla contínua requereixen una fase d'entrenament, que consisteix en aproximadament mitja hora de lectura d'un text preparat, en què sistema adquireix les dades necessàries sobre les característiques acústiques de la veu i sobre la pronúncia particular dels sons de la llengua.

La síntesi de la parla fa el camí invers al del reconeixement: la conversió de cadenes de símbols en enunciats orals. Per exemple, un sistema de síntesi per a invidents que vocalitzi el text de la pantalla d'un ordinador personal, la cadena de símbols convertida pel sintetitzador són les lletres agrupades en paraules i acompanyades per signes de puntuació.

En l'estat actual de desenvolupament tecnològic, la intel·ligibilitat de l'emissió sonora, premissa bàsica de la veu sintetitzada, es un problema ja resolt de manera pràcticament definitiva. Tot i això, encara cal solucionar la qüestió de la naturalitat de la pronunciació, és a dir, aconseguir que la veu generada per l'ordinador no soni a veu de robot. La clau per a aconseguir aquest objectiu podria ser la corba d'entonació adoptada en la generació dels enunciats, un dels aspectes de la síntesi on més s'està investigant actualment (Fernandez Rei, 1999).

3.3. Processament documental

L'àmbit del processament documental abasta una categoria molt àmplia d'aplicacions de la lingüística computacional concebudes per a l'elaboració, gestió i revisió de documents textuals.

En aquest apartat examinarem les característiques principals dels programes de verificació de la correcció lingüística dels textos, dels programes de generació automàtica de resums, dels sistemes d'extracció d'informació, dels sistemes de recuperació de la informació textual i dels programes de catalogació documental automatitzada.

3.3.1. Verificació de la correcció lingüística

La revisió automàtica de la correcció lingüística dels textos amb ajut de l'ordinador és una de les utilitats del processament de textos amb una major incidència en la qualitat dels documents produïts. Entre aquestes utilitats, les eines més utilitzades i millor considerades són els correctors ortogràfics, mentre que els programes de verificació gramatical i estilística (molts encara en fase de desenvolupament) tenen un grau d'acceptació molt menor i una eficàcia qüestionable en algunes ocasions. A continuació analitzarem el funcionament d'aquestes utilitats de verificació lingüística automatitzada, per a la qual cosa ens centrarem en la descripció formal dels seus objectius i en l'anàlisi crítica de les tècniques de revisió utilitzades (Mitton, 1996; Gómez Guinovart, 1999).

Els errors ortogràfics que es cometen durant l'escriptura d'un document amb l'ordinador es poden originar per desconeixement de la norma lingüística o per distracció; els primers reben la denominació tècnica d'*errors de competència* i els segons, d'*errors d'actuació*. Mentre que en els errors de competència la causa de l'error radica en el fet que la persona no sap com s'escriu la paraula, en els errors d'actuació la persona sí sap com s'escriu la paraula, però, per algun motiu, té un descuit o una confusió que provoca l'error.

Tipus d'errors ortogràfics

Els errors ortogràfics es poden dividir en *errors de competència* i en *errors d'actuació*.

Tot i que els errors de competència varien molt segons les persones, hi ha diversos factors lingüístics que n'afavoreixen l'aparició, com la falta de correspondència entre l'ortografia i la fonètica d'una paraula, les discrepàncies entre la normativa i l'ús o la interferència amb altres normatives, típica de les situacions de plurilingüisme. Per una altra banda, els errors d'actuació poden reflectir els errors de la parla, poden ser el resultat d'un error mecanogràfic (degut a la pulsació de tecles properes) o poden originar-se per una distracció (**problement* per *probablement*, amb el segment *-lement* posat darrere de la lletra *b* equivocada). Podem trobar classificacions detallades d'errors ortogràfics en català en les obres de Moré i altres (2005) i de Climent i altres (2003).

Tot i les seves diferents causes, i en vistes del seu tractament informàtic, la majoria d'errors ortogràfics es poden descriure mitjançant quatre mecanismes formals: inserció d'una lletra, elisió d'una lletra, substitució d'una lletra per una altra o transposició de dues lletres adjacents. A més a més, s'ha comprovat de manera empírica que, en la pràctica de l'escriptura amb ordinador es cometen molts pocs errors en la primera lletra d'una paraula. Com veurem immediatament, aquestes característiques formals dels errors ortogràfics típics de l'escriptura assistida per ordinador són en la base del disseny dels programes informàtics que serveixen per a corregir-los.

Mecanismes de descripció d'errors ortogràfics

Hi ha quatre mecanismes bàsics de descripció d'errors ortogràfics: inserció, elisió, substitució i transposició.

Els correctors ortogràfics que proporcionen els processadors de textos intenten identificar els errors ortogràfics del document i suggerir-ne la seva possible. La tècnica informàtica més habitual per a detectar aquests errors consisteix a comparar les paraules del document amb una llista de paraules correctes

emmagatzemada a l'ordinador. Aquesta llista es pot veure com un diccionari ortogràfic normatiu de la llengua que inclou les diferents formes flexives de les paraules, així com formes complexes, derivades i compostes. El corrector ortogràfic indica un error simplement quan una paraula del text no es troba en aquesta llista.

Respecte a la correcció, la tècnica informàtica clàssica aplicada en aquest cas consisteix a aplicar a la paraula que conté l'error ortogràfic els quatre mecanismes d'error esmentats de forma inversa. D'aquesta manera, quan el corrector identifica una paraula errònia en el text, busca en el diccionari les possibles formes correctes entre les paraules que comencin per la mateixa primera lletra (on generalment no es produeixen errors) i que només suposin un tipus d'error (inserció, elisió, substitució o transposició). Si la cerca resulta infructuosa, el corrector pot ampliar el seu àmbit de cerca a les paraules que comencin per una lletra diferent (**sberta* per *oberta*) o en què suposin més d'un tipus d'error (**oetra* per oberta, amb elisió i transposició).

Aquestes tècniques simples d'identificació de les paraules ortogràficament incorrectes poden fallar per diversos motius. De vegades, el programa corrector indica un error on no n'hi ha perquè la paraula buscada, tot i ser correcta, no és a la llista de paraules utilitzada pel programa. Això acostuma a passar amb els noms propis, els neologismes, els tecnicismes i les paraules poc usuals, i normalment es resol amb l'ampliació del diccionari ortogràfic normatiu que fa servir el corrector, o per l'ús de diccionaris personals o especialitzats. En altres ocasions la solució no és tan senzilla, ja que l'error ortogràfic comès dóna lloc a una altra paraula ortogràficament correcta, diferent de la que es volia escriure (per exemple, *proba* [femení de l'adjectiu *probe*, que té *probitat*, és a dir, rectitud a obrar], quan es volia escriure *prova*). Si la seqüència resultant vulnera les regles sintàctiques de la llengua, l'error podrà ser detectat per un corrector sintàctic; en canvi, si no les vulnera, la incorrecció passarà desapercebuda.

Els correctors sintàctics són els programes encarregats de reconèixer i corregir errors gramaticals presents en els enunciats d'un document. En comparació amb els correctors ortogràfics, el seu àmbit d'aplicació és molt més imprecís. Mentre que sempre es pot determinar si una seqüència de caràcters respecta o infringeix les regles ortogràfiques d'una llengua, no sempre és fàcil per l'ordinador decidir de manera automàtica si una seqüència de paraules ortogràficament correctes contenen un error sintàctic o no, ja que les indicacions recollides per les gramàtiques normatives mai no són tan exhaustives com per abastar tots els tipus d'enunciats que ha de tractar un processador de textos. La tècnica informàtica més utilitzada per a la identificació dels errors gramaticals té un enfocament casuístic i es basa en el reconeixement de certs patrons d'error prèviament establerts. Això significa que el corrector sintàctic recorrerà tot el text, analitzant-lo per tractar de detectar les seqüències de paraules que segueixin unes determinades pautes. Aquestes pautes o patrons d'error acostumen a limitar-se a un nivell gràfic i poden incorporar una suggerència

de correcció. Per exemple, un patró simple de correcció per al català podria ser (a fi de que > a fi que). Aquest patró serviria al programa corrector per a identificar la seqüència errònia (a fi de que) i per a suggerir-ne la substitució per la seqüència correcta (a fi que). La tècnica pot refinar-se introduint abreviatures i símbols en els patrons d'error. Per exemple, el corrector pot fer servir un patró com "PENSAR en VINF > PENSAR a VINF" per a detectar i corregir aquest patró per a totes les formes del verb *pensar* i qualsevol verb en infinitiu; o un patró com "si ... PLUSCSUBJ ... PLUSCSUBJ > si ... PLUSCSUBJ ... CONDCOMP" (on PLUSCSUBJ simbolitza qualsevol verb en plusquamperfet de subjuntiu, i els punts suspensius representen qualsevol seqüència de paraules dins de l'enunciat (Gómez Guinovart, 2001)).

Òbviament, els resultats d'aquesta tècnica dependran de l'amplitud i precisió dels patrons establerts pel programa. El corrector sintàctic només detectarà un error quan aquest es correspongui amb algun dels patrons previstos, i no tots els errors gramaticals són fàcilment previsibles. Per tant, la verificació sintàctica per patrons necessita complementar-se amb altres tècniques de més complexitat, com l'anàlisi sintàctica automàtica o el tractament probabilístic de la coaparició lèxica.

Un altre tipus de corrector que incorporen alguns processadors de textos són els correctors estilístics. En general, aquest tipus de correctors realitza la funció de comprovar si els trets lingüístics del document analitzat són afins o no amb les característiques atribuïdes al gènere textual al qual s'adscriu el document. Abans que el programa corrector porti a terme la revisió del document, l'usuari del sistema ha d'indicar a quina varietat estilística pertany el text examinat. D'aquesta manera, l'ordinador portarà a terme una revisió comparant les característiques del document amb els trets lingüístics establerts com a preceptius per a la categoria textual seleccionada. Normalment, aquesta categoria es pot seleccionar a partir d'uns models estilístics predefinits pel programa. Cada un d'aquests models està definit mitjançant un conjunt de trets lingüístics formals, com el nombre màxim de paraules per oració, la presència o absència de determinats girs, o el nombre màxim de sintagmes preposicionals consecutius. Tanmateix, alguns correctors permeten que l'usuari del sistema elabori els seus propis models estilístics, assignant els valors desitjats a les característiques lingüístiques proposades pel programa. Per tal que aquesta tècnica informàtica de verificació estilística assoleixi un grau considerable d'eficiència cal establir des del principi quines són les variants estilístiques o gèneres d'una llengua, i quins són els trets lingüístics que caracteritzen cada una de les variants establertes; dues exigències de difícil acompliment a les quals cal afegir la dificultat que els trets lingüístics que es fan servir en la caracterització resultin tractables informàticament.

Activitat

Verifiqueu si el processador de textos que feu servir habitualment disposa de corrector ortogràfic, gramatical i estilístic per al català.

Una de les tècniques més comunes de verificació estilística en l'escriptura assistida per ordinador consisteix a avaluar el grau de llegibilitat del text, és a dir, el grau de dificultat de comprensió del sentit del text determinat per certs

factors lingüístics quantificables, com l'extensió de les oracions, la longitud de les paraules o la quantitat de preposicions dins d'una frase. Les tècniques d'avaluació de l'elegibilitat es basen en les regularitats estadístiques que presenten els textos en aquest tipus de factors, en funció del seu grau de dificultat de lectura. Partint d'estudis lingüístics empírics, s'elaboren fórmules o equacions de llegibilitat mitjançant mètodes estadístics que, combinant els valors d'un conjunt de factors lingüístics quantificables en el text, serveixen per a predir paràmetres estimatius del seu grau de dificultat. El procediment clàssic per a l'elaboració d'aquestes fórmules fa que, en primer lloc, s'estableixi el conjunt de trets lingüístics objectivables que hipotèticament incideixen en el grau de llegibilitat d'un text; a continuació, cal obtenir mitjançant enquestes els índexs de dificultat dels textos d'un corpus, elaborat com a model de la varietat lingüística sota estudi; en tercer lloc, s'ha de calcular la correlació estadística que hi ha entre els trets estilístics preestablerts i els índexs de dificultat obtinguts empíricament; finalment, amb les dades obtingudes, s'ha d'elaborar una fórmula de llegibilitat, a mode d'equació, amb els paràmetres estilístics de més capacitat predictiva i menor grau de correlació mútua. Per tant, la fiabilitat que podem atorgar als resultats proporcionats per cada una d'aquestes fórmules dependrà de la solidesa dels seus fonaments estadístics i de l'adequació del text analitzat a la varietat estilística utilitzada com a model per a l'elaboració de la fórmula.

3.3.2. Generació automàtica de resums

La generació automàtica de resums permet presentar la informació dels documents de manera sinòptica, cosa que facilita la possibilitat de fer una avaluació visual ràpida per veure si el document s'adiu a la necessitat concreta d'informació.

La tècnica bàsica més emprada per a la generació de resums consisteix a extreure les frases considerades més significatives del text original. Les frases se seleccionen perquè inclouen certes paraules (com per exemple paraules molt freqüents en el text o paraules que apareixen en el títol) o perquè apareixen en un context determinat (com per exemple, quan la frase és la primera del document). Altres tècniques suposen la interpretació semàntica del contingut del document i la generació del text resumit. Aquestes tècniques fan servir molta informació lingüística i generalment s'apliquen a dominis temàtics restringits. Podeu trobar una bona introducció a aquestes tècniques en l'obra de Climent (2001).

3.3.3. Sistemes d'extracció d'informació

Deixant al marge la revisió automàtica de la correcció lingüística dels textos, una altra de les aplicacions de la lingüística computacional en el camp del processament documental és l'*extracció d'informació*, que consisteix a convertir textos en informació estructurada, com per exemple, en registres d'una base de dades.

L'*extracció d'informació* consisteix a convertir textos en informació estructurada.

Així, el resultat de l'extracció d'informació d'un article de diari sobre una acció terrorista podria consistir en una fitxa on constés el tipus d'incident, la data del succés, el lloc on va passar, la identificació dels seus responsables, els objectius de l'acció, les seves conseqüències i els mitjans utilitzats, com es mostra a la figura 21 (adaptació simplificada al català d'un exemple citat a l'obra de Grishman (1997)).

Figura 21. Exemple d'extracció d'informació: text i registre

19 de març. Aquest matí un comand de la guerrilla urbana d'El Salvador ha fet explotar una bomba a prop d'una central elèctrica de San Salvador. L'artefacte ha provocat estralls de diversa consideració i ha deixat una gran part de la població sense electricitat, tot i que no s'han registrat danys personals.

TIPUS D'INCIDENT	explosió
DATA	19 de març
LLOC	San Salvador
RESPONSABLES	comand de la guerrilla urbana
OBJECTIUS MATERIALS	central elèctrica
OBJECTIUS HUMANS	—
DANYS MATERIALS	estralls
DANYS PERSONALS	no
INSTRUMENT	bomba

Així doncs, el procés d'extracció de la informació comporta localitzar un conjunt de dades concretes en el text analitzat i construir una representació estructurada d'aquestes dades. La tècnica bàsica utilitzada per a la localització de les dades consisteix a identificar en el text els patrons lexicosintàctics en què es considera que es poden concretar lingüísticament les informacions buscades. Aquesta tècnica de reconeixement de patrons va precedida habitualment de l'anotació morfosintàctica de les paraules del text, i d'una anàlisi sintàctica parcial centrada en la identificació dels grups nominals i verbals dels enunciats (Grishman, 1997; Appelt, 1999).

A diferència de la comprensió del llenguatge natural (Allen, 1995), l'extracció d'informació no pretén representar tota la informació d'un text, sinó únicament la informació seleccionada, amb la finalitat d'oferir una via eficient de consulta als grans volums de dades escrites en llenguatge natural en les notícies dels diaris, en els informes mèdics hospitalaris o en les sentències judicials.

3.3.4. Sistemes de recuperació d'informació

La recuperació d'informació textual és una tecnologia orientada a la gestió de bases de dades documentals (Strzalkowski, 1999). L'objectiu és seleccionar els documents més rellevants en relació amb uns determinats requisits d'informació expressats en una consulta.

La *recuperació d'informació* té com a objectiu seleccionar els documents més rellevants en relació amb uns requisits d'informació expressats en una consulta.

Les consultes poden combinar els termes cercats mitjançant operadors lògics i condicions de proximitat. Per a la selecció dels documents es fan servir dues llistes de paraules: una de formada per totes les paraules que apareixen en els documents i que conté també la seva localització en els textos, i l'altra amb les paraules considerades irrellevants per a les cerques documentals (Codina Bonilla, 1993). La consulta d'una base de dades textual mitjançant un sistema de recuperació d'informació ofereix com a resultat una llista de documents de la base de dades que el sistema considera rellevants per a satisfer la consulta. Per exemple, el resultat de buscar en la base de dades de notícies de l'agència EFE les empreses que van firmar contractes relacionats amb les telecomunicacions durant l'any passat consistiria en una llista de notícies que s'hauria de repassar visualment per a extreure la informació desitjada. En contrast, un sistema d'extracció d'informació oferiria directament la llista de les empreses.

3.3.5. Catalogació documental automatitzada

Les tècniques de catalogació documental intenten determinar automàticament el contingut general dels textos analitzats, per poder classificar-los dins d'una tipologia semàntica preestablerta.

Per exemple, un text bancari es pot catalogar com a "dèbit per domiciliació" i un altre com a "contracte de compte". L'índex bibliogràfic que s'obté de la classificació documental constitueix una via d'accés simple i directa a la informació continguda en els textos.

3.4. Traducció automàtica

La traducció automàtica és una de les aplicacions de la lingüística computacional de més gran complexitat i un dels desenvolupaments de més interès

per al públic no especialista. En sentit ampli, el camp de la traducció automàtica inclou tot el conjunt d'eines informàtiques dissenyades per a la seva incorporació en el procés de la traducció humana. Les aplicacions que formen part d'aquest conjunt poden agrupar-se segons criteris diferents: el nombre de parells de llengües entre els quals el sistema tradueix (sistemes bilingües o plurilingües), si tradueix els parells de llengües en una única direcció o també pot fer la traducció inversa (sistemes unidireccionals o bidireccionals), si està limitat en un àmbit lingüístic o en una àrea temàtica concreta o en un determinat tipus de llengua simplificada (traducció de llenguatge en general o traducció de subllenguatges), segons la metodologia de traducció aplicada (traducció directa, per transferència o mitjançant interlingua), o segons el grau d'automatització de la traducció (traducció automàtica o traducció assistida per ordinador) (Hutchins i Somers, 1992; Whitelock i Kilby, 1995).

El terme traducció automàtica, en sentit estricte, es refereix als programes de traducció que no requereixen intervenció humana per a realitzar la seva tasca. Fins ara aquest tipus d'aplicacions només ofereixen un grau de fiabilitat acceptable en la traducció de subllenguatges, particularment en dominis de coneixement molt restringits (llenguatges sectorials) o quan el text de partida està escrit seguint unes normes molt estrictes orientades a la simplificació del lèxic i sintaxi (llenguatges controlats). Per exemple, el programa TAUM-MÉTEO, que fa servir intensivament des del 1977 pel Departament de Medi Ambient del Canadà, tradueix els parts meteorològics de l'anglès al francès sense que es precisi pràcticament revisió humana.

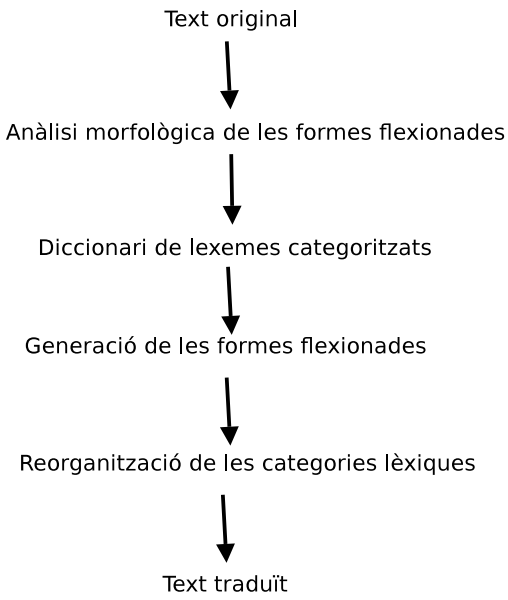
Dins de la categoria de traducció assistida per ordinador cal fer distinció entre la traducció semiautomàtica (amb intervenció humana) i la traducció (humana) amb ajut de l'ordinador. Els programes informàtics de traducció semiautomàtica ofereixen una traducció esborrany del text original que ha de ser revisada en profunditat per poder aconseguir una qualitat de traducció similar a la d'un traductor professional. En l'àmbit de la informàtica personal, els programes d'aquest tipus més populars són els comercialitzats amb diverses denominacions per l'empresa Globalink. Per estacions de treball el sistema més emprat és SYSTRAN, adoptat des del 1981 per la Comissió de les Comunitats Europees per a les seves traduccions.

Els programes de traducció amb ajuda de l'ordinador estan concebuts per a col·laborar com a assistents en la traducció humana d'un text. Per exemple, els entorns integrats de treball amb memòria de traducció, com el TranslationManager d'IBM, Trados o Déjà Vu integren en un únic producte informàtic un processador de textos especialment dissenyat per a traduir, un conjunt de diccionaris bilingües, eines de gestió de les bases de dades lèxiques i una memòria de traducció. Una memòria de traducció és una base de dades on s'emmagatzemen la versió original i traduïda de cada una de les oracions que es van traduint. Quan es tradueix una frase, el programa detecta automàticament si aquesta mateixa frase o una de semblant ja ha estat traduïda prèviament,

amb l'objectiu de reaprofitar la traducció sense haver-la de tornar a escriure, però podent fer les modificacions que siguin necessàries.

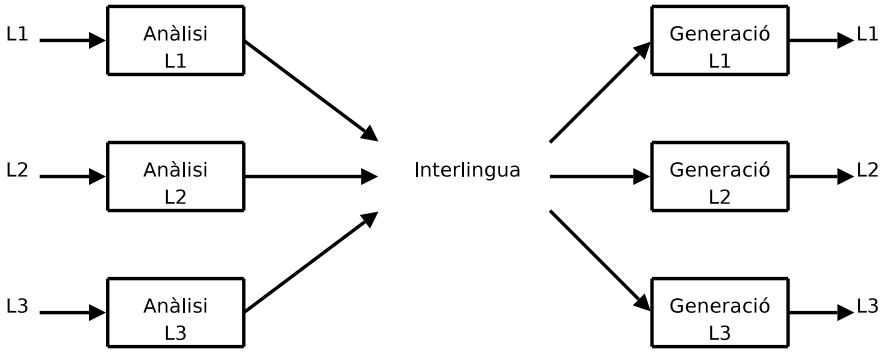
L'estratègia o model de traducció aplicada pels programes de traducció automàtica o semiautomàtica pot ser directa, per transferència o mitjançant interlingua. En la *traducció directa*, el processament del text original produeix directament el text traduït, sense que hi hagi cap tipus d'anàlisi sintàctica o semàntica intermèdia. En la seva versió més simple, es pot tractar d'una traducció paraula per paraula, on la traducció es porta a terme substituint cada paraula de l'original per la paraula corresponent en el diccionari bilingüe del sistema. Com a complement acostuma a fer-se servir un processador morfològic de les paraules del text original que permeti utilitzar un diccionari més reduït de formes no flexionades, o algun tipus de reorganització senzilla de les categories lèxiques del text traduït (com per exemple, en la traducció entre el català i l'anglès, invertir l'ordre de les paraules quan es detecti una seqüència formada per un nom seguit d'un adjectiu) (figura 22).

Figura 22. Estratègia de traducció directa ampliada



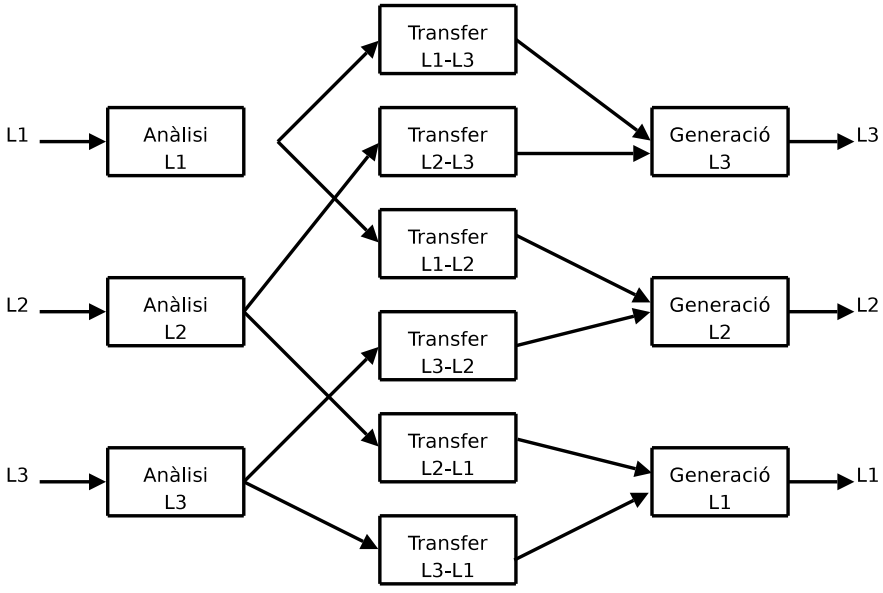
En la traducció mitjançant *interlingua* s'analitza el text original per construir una representació intermèdia a partir de la qual es genera directament el text de la traducció. La representació intermèdia és una formalització del significat conceptual del text, per la qual cosa constitueix una representació abstracta tant del text original com del text traduït. El terme *interlingua* al·ludeix al fet que aquesta representació semàntica pretén ser neutral respecte a la llengua i fins i tot, en alguns casos, pretén ser universal. L'estratègia és molt adequada per als sistemes plurilingües ja que, un cop definida la interlingua, només exigeix dos mòduls de programació per a cada llengua L incorporada bidireccionalment al sistema: un mòdul d'anàlisi (o traducció de L a la interlingua) i un de generació (o traducció de la interlingua a L) (figura 23).

Figura 23. Estratègia de traducció mitjançant interlingua



En el mètode de *traducció per transferència*, la traducció es realitza en tres fases: la fase d'anàlisi del text original, que produeix com a resultat una representació sintacticosemàntica depenent de la llengua analitzada; la fase de transferència, en què se substitueixen les paraules de la llengua original per les paraules de la llengua de la traducció i es converteix l'estructura sintacticosemàntica pròpia de la llengua original en una estructura equivalent en la llengua de la traducció, i la fase de generació, en què es transforma la representació sintacticosemàntica fruit de la transferència a un text en la llengua de la traducció. Respecte al model d'interlingua, aquesta estratègia té l'avantatge que els seus mòduls d'anàlisi i generació són menys complexos, ja que treballen amb representacions molt properes a les llengües representades. Per contra, el seu principal defecte és el nombre de mòduls necessaris per a construir un sistema plurilingüe: per a traduir bidireccionalment entre n llengües, cal construir un total de $n^2 + n$ mòduls, davant els $2n$ mòduls necessaris en l'aproximació per interlingua. Per exemple, en un sistema trilingüe per transferència es necessiten un total de 12 mòduls (6 de transferència, 3 d'anàlisi i 3 de generació), davant els 6 necessaris en un sistema d'interlingua (3 d'anàlisi i 3 de generació) (figura 24).

Figura 24. Estratègia de traducció per transferència



Per últim cal presentar els anomenats *sistemes de traducció estadístics* i els *sistemes de traducció basats en exemples*. Aquestes dues tècniques de traducció automàtica es basen principalment en l'explotació de corpus paral·lels. El que intenten fer és traduir noves frases a partir de les frases originals i traduïdes que es troben en el corpus. La probabilitat que hàgim de traduir novament una frase present en el corpus no és gaire elevada, però a partir dels exemples que es trobem en el corpus paral·lel, i aplicant diverses tècniques, podrem intentar donar la traducció de noves frases.

Per l'explicació de la tècnica de *traducció automàtica estadística* considerarem com a exemple la traducció d'un text del català a l'anglès (adaptat de l'obra de Brown i altres (1993)). Donada una frase en català C, ens imaginem que aquesta prové d'una frase en anglès, que anomenarem A. Per a obtenir la frase en català, la frase en anglès s'ha transmès a través d'un canal de comunicació sorollós, és a dir, que introdueix una distorsió a la frase. Aquest canal té la curiosa propietat que les frases en anglès que es transmeten es reben en català. La presumpció principal dels sistemes de traducció automàtica estadístics és que les característiques d'aquest canal es poden obtenir experimentalment i que es poden expressar matemàticament.

Aquest formalisme es pot fer servir per a obtenir traduccions del català a l'anglès de la manera següent: $Pr(a|c)$ és la probabilitat que "a" sigui la frase en anglès que doni com a resultat la frase en català "f". Donada la frase en català "c", el problema de la traducció automàtica es redueix a trobar la frase en anglès que maximitzi $Pr(a|c)$, és a dir, trobar:

$$\hat{a} = argmax_e Pr(a|c)$$

Aplicant el teorema de Bayes obtenim:

$$\hat{a} = argmax_e Pr(a|c) = argmax_e Pr(c|a).Pr(a)$$

$Pr(a|c)$ és el model de traducció, que modela la probabilitat que "c" provingui del canal quan "a" és a l'entrada del canal. El domini d'aquesta funció són tots els parells <c,a> de cadenes de paraules en català i en anglès. $Pr(a)$ és el model de llengua i modela la probabilitat que "a" estigui a l'entrada del canal. Cada un d'aquests factors (el model de traducció i el model de llengua) independentment produeixen un índex per a una traducció anglesa candidata "a". El model de traducció ens proporciona un índex que ens indica en quin grau les paraules de "a" expressen el que diu "c", i el model de llengua ens diu en quin grau "a" és una frase gramatical. La traducció seleccionada serà aquella que maximitzi aquest producte. Els models $Pr(c|a)$ i $Pr(e)$ s'obtenen automàticament a partir de corpus paral·lels.

La *traducció automàtica basada en exemples* pretén realitzar la traducció d'una determinada frase a partir d'un conjunt d'exemples de frases originals amb les

seves corresponents traduccions (Way, 2001). En aquest sentit aquesta tècnica fa servir bàsicament el mateix recurs que la traducció automàtica estadística, és a dir, un corpus paral·lel. La diferència principal radica en el fet que la traducció automàtica basada en exemples no calcula una sèrie de paràmetres estadístics, sinó que classifica els exemples de traducció i els intenta generalitzar, de manera que es puguin fer servir per a traduir altres frases. Posem, per exemple, que volem traduir la frase:

El Sr. Martí viatjarà a Madrid

i disposem de l'exemple

El Sr. Anglada viatjarà a Barcelona

amb la seva traducció

El Sr. Anglada viajará a Barcelona.

Si generalitzem aquest exemple de la manera següent:

El Sr. COGNOM viatjarà a CIUTAT
El Sr. COGNOM viajará a CIUTAT

i sabem que Martí és un COGNOM i Madrid una CIUTAT, podrem traduir fàcilment la frase inicial.

El pas de generalització no és sempre senzill i requereix cert coneixent lingüístic (en l'exemple, una llista de cognoms i de ciutats). Hi ha diferents variants de traducció automàtica basada en exemples (Hutchins, 2005).

4. Informàtica aplicada a la traducció

La denominació d'*informàtica aplicada a la lingüística* (o simplement *lingüística informàtica*) engloba, en la seva accepció més general, tota la varietat d'estudis lingüístics que fan servir eines informàtiques per a la seva investigació i, de manera més específica, aquells en què l'aplicació de la lingüística té una influència més notòria per a l'obtenció dels resultats.

En aquesta línia examinarem de forma resumida dos exemples il·lustratius de la investigació en aquest camp, centrant la nostra atenció en la *lingüística comparada diacrònica* i en la *lingüística de corpus*. Altres línies d'investigació destacables de la lingüística informàtica són la *informàtica aplicada a la lingüística antropològica* (Antworth i Valentine, 1998), la *informàtica aplicada a la lexicografia* (Boguraev i Briscoe, 1989; Ooi, 1998; Martí i altres, 1998; Álvarez Lugrís, 1997; Pérez Hernández i altres, 1999) i la *informàtica aplicada a la sociolingüística* (Moreno Fernández, 1994; Lorenzo Suárez, 1999; Ramallo, 1999).

4.1. Lingüística de corpus

La *lingüística de corpus* és una disciplina dedicada a l'estudi empíric del llenguatge a partir de les dades que proporcionen els corpus lingüístics (Badia, 1996; Llisterri, 1996; McEnery i Wilson, 1999; Pérez Guerra, 1998).

Un corpus lingüístic consisteix en una col·lecció de textos reals escrits o parlats, emmagatzemats en suport informàtic, compilats per a portar a terme algun tipus d'investigació o aplicació lingüística, i representatius d'una varietat lingüística determinada.

La selecció dels textos que formen part d'un corpus depèn dels objectius del corpus. Els corpus generals pretenen recollir totes les varietats i registres d'una llengua, amb l'objectiu d'assolir la màxima representativitat lingüística. Mentre que els corpus especialitzats (monolingües o plurilingües) acostumen a centrar-se en una varietat lingüística concreta (per exemple, el Corpus Legebiduna inclou textos administratius i jurídics bilingües publicats en euskera i castellà per les institucions del País Basc) (Abaitua i altres, 1997). Quan els textos plurilingües recopilats són versions traduïdes dels mateixos documents

es parla de corpus de textos paral·lels; en els corpus de textos alineats, a més, estan identificades les equivalències de traducció entre els segments (paraules, frase o unitats de traducció) de cada una de les versions traduïdes (Hallebeek, 1999).

Els corpus sense anotar, és a dir, els que contenen exclusivament les paraules i els signes de puntuació dels textos originals, tenen una utilitat bastant limitada per a la investigació en lingüística de corpus. Així, en un corpus sense anotar seria complicat trobar automàticament (i sense cap tipus d'anàlisi prèvia) els sintagmes nominals, els tonemes descendents o les oracions en què el subjecte no està en posició inicial. Per facilitar aquestes cerques i altres de semblants, els corpus anotats afegeixen diversos tipus d'informació lingüística que es pot elaborar de manera manual, semiautomàtica o completament automàtica.

Actualment, l'anotació morfosintàctica, és a dir, el procés d'assignar una marca de categoria gramatical a cada paraula d'un text, es pot portar a terme amb un alt grau d'automatització. Els programes informàtics de marcatge gramatical automàtic analitzen les seqüències de paraules per produir les etiquetes morfosintàctiques corresponents. La llista d'etiquetes dependrà de les característiques lingüístiques del corpus, dels objectius de la seva anotació, dels límits del programa i dels pressupòsits teòrics aplicats (Garside i altres, 1997). La figura 25 mostra un exemple d'anotació gramatical real (lleugerament modificada) del Corpus LOB (apud McEnery i Wilson (1999)).

Figura 25. Exemple de marcatge morfosintàctic

Joanna stubbed out her cigarette with unnecessary fierceness. Her lovely eyes were defiant above cheeks whose colour had deepened at Noreen's remark.
Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN unnecessary_JJ fierceness_NN ._. Her_PP$ lovely_JJ eyes_NNS were_BED defiant_JJ above_IN cheeks_NNS whose_WP$ colour_NN had_HVD deepened_VBN at_IN'Noreen s_NP$ remark_NN ._.

Els etiquetadors morfosintàctics automàtics actuals poden fer servir regles lingüístiques o models probabilístics com a mecanisme d'assignació d'etiquetes. També hi ha sistemes híbrids que combinen aquestes dues estratègies en funció de les seves necessitats. La majoria dels etiquetadors probabilístics assignen les etiquetes fent servir la informació morfològica i conceptual de tipus estadístic derivada automàticament a partir d'un corpus etiquetat prèviament. L'etiquetador fonamentarà la seva actuació en dues estimacions estadístiques: la probabilitat lèxica, basada en la freqüència relativa amb què una paraula rep una etiqueta determinada en el corpus inicial; i la probabilitat contextual, basada en la freqüència relativa amb què una determinada etiqueta apareix abans de les *n* etiquetes següents i/o després de les *n* etiquetes anteriors en el corpus inicial. El valor de *n* dependrà dels etiquetadors, però sembla que la millor eficiència s'obté assignant a *n* el valor 1 o 2.

Adreça recomanada

Podeu provar l'etiquetador *Freeling* accedint a www.lsi.upc.edu/~nlp/freeling. A part de provar-lo en línia podeu descarregar-lo i instal·lar-lo al vostre ordinador.

El marcatge de les paraules desconegudes (és a dir, les que no apareixen en el corpus inicial) és un problema important dels etiquetadors probabilístics. La

solució més habitual consisteix a etiquetar la paraula desconeguda en funció del seu context, ponderant aquest factor contextual amb altres factor deduïbles a partir de la grafia de la paraula. Per exemple, la presència d'una majúscula inicial faria augmentar la probabilitat d'assignació de l'etiqueta corresponent als noms propis, i si la paraula acaba en *-ció* incrementaria la probabilitat de rebre una etiqueta de substantiu.

Els corpus anotats morfosintàcticament són molt útils per a la investigació lingüística i per al desenvolupament d'aplicacions de processament del llenguatge natural, ja que proporcionen textos amb paraules no ambigües respecte a la seva categoria gramatical.

Un corpus anotat amb informació sintàctica, és a dir, un corpus analitzat sintàcticament i etiquetat amb aquesta informació resulta encara més valuós. Tot i això, hi ha pocs corpus extensos anotats sintàcticament, ja que s'han d'elaborar de manera manual o semiautomàtica, atès que l'anàlisi sintàctica automàtica encara presenta nombrosos problemes per a la seva automatització completa. A la figura 26 es pot observar un exemple de marcatge sintàctica del *Lancaster Parsed Corpus* (apud Pérez Guerra (1998)).

Figura 26. Exemple de marcatge sintàctic

I can n't make a club pay a player so much a week. [S[Na la I_PP1A Na] [V can_MD n't_XNOT make_VB V][N a_AT club_MM N][Tb[V pay_VB V] [N a_AT player_NN N][N[D so_QL much_AP D][N a_AT week_NN N]N]Tb]._. S]

Ocasionalment també es poden trobar altres tipus d'anotació lingüística en els corpus, com pot ser l'*anotació prosòdica* (indicacions sobre pauses en la pronúncia, corves d'entonació, grups fònics, indicacions d'èmfasi, etc.), l'*anotació semàntica* (indicacions sobre característiques semàntiques dels elements lèxics o sobre si pertanyen a un determinat camp semàntic), l'*anotació discursiva* (com per exemple indicacions sobre la referència dels pronoms) i l'*anotació dels fenòmens propis de la conversa* (com per exemple, indicacions sobre els torns de parla).

Pel que fa al format de marcatge, hi ha moltes maneres de representar la informació lingüística en els textos. En general qualsevol dels formats de marcatge respecten la norma que totes les etiquetes es puguin suprimir amb facilitat. Les anotacions de la figura 27, per exemple, poden ser eliminades fàcilment borrant les seqüències de caràcters situats entre un guió i un espai.

Un altre format d'anotació cada cop més habitual en el marcatge de corpus és el format SGML (sigla de Standard Generalised Markup Language) i, en particular, un subconjunt de l'SGML conegut com a format TEI (*Text Encoding Initiative*) (Sperberg-McQueen i Burnard, 1994). Sense entrar en gaires detalls, ja que una presentació exhaustiva de les normes TEI excediria els límits d'aquest mòdul, cal saber que aquestes directrius proporcionen un conjunt normalitzat

d'etiquetes i abreviatures per a la representació descriptiva i estructurada de la informació textual continguda en els corpus. Les etiquetes TEI s'han d'escriure entre < i >. Els documents en format TEI es divideixen en dues parts: la capçalera i el mateix text anotat. La capçalera conté les informacions bibliogràfiques i de codificació relatives al document electrònic: autoria, data de publicació, títol. Edició, origen de la versió electrònica, observacions sobre el format d'anotació utilitzat, etc. Per una altra banda, el text s'estructura en tres seccions: els preliminars (dades de la primera plana, prefaci, dedicatòria, índex, etc.), el cos (subdividit jeràrquicament en diversos elements) i la part final (apèndixs, notes, bibliografia, glossari, etc.). Com a exemple, la figura 27 conté un exemple d'anotació textual amb TEI d'un fragment del conte "Tinta china" de Gonzalo Navaza.

Figura 27. Exemple de marcatge TEI

```
<tei.2>
<teiheader> <filedesc> <titlestmt> <title> Fragmentos del cuento "Tinta china", de Gonzalo Navaza: un ejemplo de codificaci&oacute;n TEI </title> <respstmt> <resp> preparado por </resp> <name>Xavier G&oacute;mez Guinovart </name> </respstmt> </titlestmt> <publicationstmt> <publisher> Edici&oacute;ns Generales de Galicia </publisher> </publicationstmt> <sourcedesc> <bibl> <author> Gonzalo Navaza </author> <title type=ïtem» Tinta china </title> <title> Errores y T&aacute;natos </title> <imprint> <publisher> Edici&oacute;ns Generales de Galicia </publisher> <pubplace> Vigo </pubplace> <date> 1996 </date> </imprint> <extent> pp. 43-56 </extent> </bibl> </sourcedesc> </filedesc> </teiheader>
<text> <front> <titlepage> <doctitle> <titlepart> Tinta china </titlepart> </doctitle> </titlepage> <div1 type="dedication» <pb> <p> A Arthur M. Morgan, <foreign> in memoriam </foreign> . </p> </div1> </front>
<body> <p> En <date value="09-1993» septiembre de 1993 </date>, durante los preparativos de la celebraci&oacute;n del congreso del <foreign>PEN Club</foreign> Internacional en Santiago de Compostela, los pobres chicos y menos chicos aspirantes a ingresar en tan ilustre asociaci&oacute;n de escritores recibimos de la organizaci&oacute;n el encargo de recoger en el aeropuerto alg&uacute;ns de los asistentes y conducirlos incluso el hotel. Trat&aacute;base en general de escritores de segunda fila, pues los pesos pesados &iacute;an ser atendidos directamente polos organizadores. El encargo inclu&iacute;la la posibilidad de hacer de cicerone por las r&uacute;las y monumentos de la ciudad y obligaba a estar siempre dispuesto para resolver cualquier #problema
que pudiera present&aacute;rselles &oacute;s convidados. </p> <note resp="XGG» texto elidido (dos par&aacute;grafos) </note> <pb n="46» <p> De aquella despedida quiero evocar tam&eacute;n, si se me permite, otro detalle. Cuando ya se habían retirado los fot&oacute;grafos y el resto de los acompa&ntilde;antes, m&iacute;ster Morgan coleunos &aacute; parte a Marcelo Cardalda y la mí y, obrig&aacute;ndonos a colocar las palmas de las manos como en el juramento de los tres mosqueteros, chilló en franc&eacute;s <q rend=«< »» <foreign> Un pour tous, tous pour uno! </foreign> </q> , con la s&uacute;la voz de b&uacute;falo y su curioso acento. Luego los dio un abrazo efusivo y antes de perderse polo corredor de embarque los entregó un pequeño obsequio: la Carralda <title rend=ïtalic» <foreign> A Touch of Class </foreign> </title> , en una edici&oacute;n ilustrada por Bacon, y la mí los <title rend=ïtalic» <foreign> Forgettable Tales </foreign> </title> , en la edici&oacute;n de Planet, encuadernada en piel. </p> <note resp="XGG» resto de texto elidido </note> </body> </text> </tei.2>
```

4.2. La lingüística històrica computacional

La lingüística històrica estudia els canvis que es produeixen en les llengües en una dimensió temporal o diacrònica. Un dels seus objectius consisteix a reconstruir deductivament els estadis hipotètics anteriors d'una llengua a partir de les formes lingüístiques posteriors testimoniades.

Quan aquestes formes pertanyen a diferents varietats lingüístiques, presumiblement relacionades entre si, la reconstrucció de la protolingua, és a dir, de la llengua comú hipotètica de què es van originar, es fa mitjançant el mètode comparatiu. El procés de reconstrucció lingüística comença amb la identificació dels conjunts de paraules derivades d'un ètim en les diverses llengües (com per exemple, l'anglès *father*, el grec *pater* i el sànscrit *piter*) i, a partir dels grups de cognats, es reconstrueixen els ètims (com per exemple ***pAter*) i es formulen les regles diacròniques que descriuen els canvis fonètics observats (en anglès *p > f).

El programa PHONO*permet verificar els efectes de les teories postulades si se li subministra un conjunt ordenat de regles fonològiques i una descripció dels trets distintius de les vocals i consonants que componen l'alfabet. A partir d'un ètim de la protollengua, PHONO generarà automàticament la cadena derivada prevista en la teoria, manipulant les matrius de trets d'acord amb les regles preestablertes (figura 28).

*http://mypage.siu.edu/lhartman

Figura 28. Derivació de *mensa* del llatí al castellà *mesa* amb PHONO

```
ETYMON -->                    mensa
  HOMORGANIC:            =>   ménSa  -ante-dist
  PRENASAL_LONG:         =>   mén Sa  +long/-ante-dist
  NS_S:                  =>   méSa  +long
  UNLONG:                =>   méSa
  VOICING:               =>   méZa
  UNVOICE:               =>   méSa
```

Altres programes informàtics, com COGNATE**(Guy, 1994) i WORDSURV***, ofereixen la possibilitat de calcular el grau de parentiu entre paraules de diferents llengües, basant-se exclusivament en la probabilitat estadística de les seves correspondències fonètiques (figura 29).

**ftp://garbo.uwasa.fi/pc/linguistics/cognate.zip
***ftp://ftp.sil.org/software/dos/wrdsrv25.zip

Figura 29. Probabilitats de cognació entre *apple* i *apfle* amb COGNATE

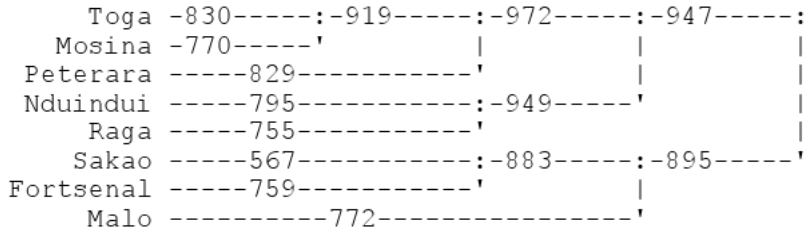
```
    a   p   p   l   e
a  93  32  32  78  68
p  21  83  83  66  14
f  34  86  86  52  70
e  49  63  63  30  91
l  44  54  54  97  41

apfel/apple  (word pair #3, pass #1)
I am 39% sure that they ARE related.
I allowed only for matches >= 50.
The best I found were: apfel
                       app le
```

Finalment, hi ha també utilitats com RECONSTRUCTION ENGINE (Lowe i Mazaudon, 1994) per a la reconstrucció de les formes d'una protollengua a partir de les dades lingüístiques de qualsevol grup de llengües del món, i com GLOTTO**** , que és capaç d'elaborar automàticament arbres genealògics de famílies lingüístiques a partir de llistes de paraules i de la seva anàlisi lexi-

****ftp://garbo.uwasa.fi/pc/linguistics/glotto02.zip

coestadístic. A la figura 30 s'ofereix un exemple dels resultats de GLOTO, a partir de dades de vuit llengües austronèsiques de Vanuatu, on les quantitats expressen la proporció de vocabulari (per cada mil paraules) que es manté en cada divisió de la família (com per exemple, el sakao i el fortsenal mantindries respectivament el 56,7% i el 75,9% del lèxic del seu avantpassat comú que, per la seva banda, conservaria un 88,3% de la seva llengua antecessora).

Figura 30. Arbre genealògic lexicoestadístic elaborat per GLOTTO



Resum

En aquest mòdul hem procurat presentar, des d'una perspectiva didàctica i global, algunes de les línies de treball més destacables de la lingüística computacional. Amb aquest objectiu, en l'apartat "Àmbit de la lingüística computacional" hem intentat delimitar el camp d'estudi d'aquesta disciplina i d'establir les seves relacions amb les altres àrees de la lingüística i amb la societat. A continuació, en l'apartat "Models i formalismes lingüístics", hem presentat el camp de treball de la lingüística computacional teòrica, atenent particularment els models lingüístics simbòlics i els formalismes lingüístics d'unificació. En l'apartat "Aplicacions de la lingüística computacional" hem presentat l'estat de les tecnologies de la llengua en els camps de la comprensió i generació del llenguatge, del reconeixement i de la síntesi de la parla, de l'extracció d'informació textual i de la traducció automàtica. Per últim, en l'apartat "Informàtica aplicada a la traducció", hem examinat alguns usos de la informàtica en la investigació lingüística, centrant-nos en la seva aplicació a la lingüística històrica i, de manera especial, a l'anàlisi textual de corpus lingüístics. Mitjançant la lectura atenta d'aquest mòdul i de les referències bibliogràfiques proposades, les persones interessades podran obtenir una visió panoràmica general d'un dels camps de la lingüística amb més possibilitats d'investigació i desenvolupament, i amb una major incidència social.

Bibliografia

Abaitua, J.; Casillas, A.; Martínez, R. (1997). «Segmentación de corpus paralelos para memorias de traducción». A: *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, (21): pàgs. 17–30.

Allen, J. (1995). *Natural Language Understanding*. 2a edició. Redwood: Benjamin/Cummings. *Resum*: Monografia acadèmica dedicada a l'estudi dels sistemes de comprensió del llenguatge. Analitza amb profunditat, des d'una perspectiva computacional, els problemes sintàctics, semàntics i pragmàtics implicats en la conversió d'un text en una representació del seu significat.

Alonge, A.; Calzolari, N.; Vossen, P.; Bloksma, L.; Castellón, I.; Martí, M. A.; Peters, W. (1998). «The Linguistic Design of the EuroWordNet Database». A: *Computers and the Humanities*, (32): pàgs. 91–115.

Álvarez Lugrís, A. (1997). «Técnicas de representación de la lexicografía plurilingue». A: *Revista Española de Lingüística Aplicada*, volum 1. ISSN 0213-2028.

Antworth, E.; Valentine, R. (1998). *Software for Doing Field Linguistics*, volum Using Computers in Linguistics: a Practical Guide, (pàgs. 170–196). Londres: Routledge.

Appelt, D. (1999). «Introduction to Information Extraction». A: *AI Communications*, (12): pàgs. 161–172.

Badia, T. (1996). «El processament computacional de corpus: tècniques automàtiques d'anàlisi morfològica i sintàctica». A: «Actes del 1r i 2n col·loquis lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)», (eds.) Payrató, L.; Boix, E.; LLoret, M. R.; Lorente, M. (pàgs. 217–254). Barcelona: Promociones y Publicaciones Universitarias. ISBN 8447705935.

Balari Ravera, S. (1999). «Formalismos gramaticales de unificación y procesamiento basado en restricciones». A: *Revista Española de Lingüística Aplicada*, volum 1. ISSN 0213-2028.

Bird, S. (1995). *Computational Phonology: A Constraint-Based Approach*. Cambridge: Cambridge University Press.

Boguraev, B.; Briscoe, T. (1989). *Computational Lexicography for Natural Language Processing*. Londres: Longman.

Bresnan, J. (1999). *Lexical-Functional Syntax*. Oxford: Blackwell.

Brown, P. F.; Dell Pietra, S. A.; Dell Pietra, V. J.; Mercer, R. L. (1993). «The Mathematics of Statistical Machine Translation». A: *Computational Linguistics*, volum 19: pàgs. 263–311.

Carpenter, B.; Penn, G. (1997). *ALE: The Attribute Logic Engine*. Pittsburgh: Carnegie Mellon University.

Charniak, E. (ed.) (1993). *Statistical Language Learning*. Cambridge: The MIT Press.

Climent, S. (2001). «Sistemes de resum automàtic de documents». A: *Digithum*, (3). ISSN 1575-2275.

Climent, S.; Moré, J.; Oliver, A.; Salvatierra, M.; Sánchez, I.; Taulé, M.; Vallmanya, L. (2003). «Bilingual Newsgroups in Catalonia: a Challenge for Machine Translation». A: *Journal of Computer Mediated Communication*, volum 9(1).

Codina Bonilla, L. (1993). *Sistemes d'informació documental*. Barcelona: Pòrtic.

Dutoit, T. (1997). *An Introduction to Text-to-speech Synthesis*. Kluwer Academic Publishers. *Resum*: Manual sobre els sistemes de conversió de text a veu dirigit a estudiants d'aquesta matèria en lingüística computacional o enginyeria de telecomunicacions. Presenta les tècniques de processament del llenguatge natural i del processament del senyal sonor que es fan servir actualment en la síntesi de la parla.

Evans, R.; Gazdar, G. (1996). «DATR: A Language for Lexical Knowledge Representation». A: *Computational Linguistics*, (22): pàgs. 167–216.

Fernandez Rei, E. (1999). «Tecnologías del habla y síntesis de voz en gallego». A: *Revista Española de Lingüística Aplicada*, volum I: pàgs. 103–116.

Garside, R.; Klein, E.; McEnery, T. (eds.) (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Oxford: Blackwell.

Gazdar, G.; Klein, E.; Pullum, G.; Sag, I. (1985). *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

Gazdar, G.; Mellish, C. (1989). *Natural Language Processing in Prolog: An Introduction to Computational Linguistics*. Wokinham: Addison-Wesley. *Resum*: Manual de lingüística computacional concebut com a llibre de text d'orientació pràctica. Conté una extensa selecció d'exemples i exercicis de programació en Prolog que il·lustren els conceptes i les tècniques de processament sintàctic, semàntic i pragmàtic presentats en el text.

Gómez Guinovart, J. (1999). *La escritura asistida por ordenador*. Vigo: Universidade de Vigo (Servicio de Publicacións).

Gómez Guinovart, J. (2001). *Recursos s'ajut a l'edició*, volum Llengua Catalana IV: les tecnologies del llenguatge. Barcelona: Universitat Oberta de Catalunya.

Gómez Guinovart, X. (2000). «Lingüística computacional». A: «Manual de ciencias da linguaxe», (eds.) Ramallo, F.; Rei Doval, G.; Rodríguez, X. P., capítol Lingüística Computacional, (pàgs. 221–268). Xerais.

Grishman, R. (1997). *Information Extraction: Techniques and Challenges*, volum Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, (pàgs. 10–27). Berlin: Springer-Verlag.

Guy, J. (1994). «An Algorithm for Identifying Cognates in Bilingual Word-lists and its Applicability to Machine Translation». A: *Journal of Quantitative Linguistics*, (1): pàgs. 35–42.

Hallebeek, J. (1999). «El corpus paralelo». A: *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, (24): pàgs. 49–55.

Hutchins, J. (2005). «Towards a definition of example-based machine translation». A: «Proceedings of MT Summint X Workshop: Second workshop on example-based machine translation», (pàgs. 63–70). Pukhet (Thailand).

Hutchins, J.; Somers, H. (1992). *An Introduction to Machine Translation*. Londres: Academic Press. Traducció al castellà: *Introducción a la traducción automática*. Visor. Madrid. 1995, *Resum*: Es tracta d'un complet manual sobre traducció automàtica. Inclou una àmplia introducció als fonaments lingüístics i computacionals de la traducció automàtica i una descripció minuciosa de les característiques de disseny dels diferents sistemes.

Kay, M. (1982). *Parsing in Functional Unification Grammar*, volum Natural Language Parsing. Cambridge: Cambridge University Press.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production*. Helsinki: University of Helsinki.

Llamas, C.; Cardeñoso, V. (1997). *Reconocimiento automático del habla: técnicas y aplicación*. Valladolid: Universidad de Valladolid.

Llisterri, J. (1996). «Els corpus lingüístics orals». A: «Actes del 1r i 2n col·loquis lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2)», (eds.) Payrató, L.; Boix, E.; LLoret, M. R.; Lorente, M., (pàgs. 27–70). Barcelona: Promociones y Publicaciones Universitarias.

Lorenzo Suárez, A. (1999). «Sociolingüística cualitativa y lingüística informática». A: *Revista Española de Lingüística Aplicada*, (1): pàgs. 247–261. ISSN 0213-2028.

Lowe, J.; Mazaudon, M. (1994). «The Reconstruction Engine: a Computer Implementation of the Comparative Method». A: *Computational Linguistics*, (20): pàgs. 381–417.

Martí, M. A.; Castellón, I.; Fernández, A. (1998). *Extracción de información de corpus diccionariales*, volum Lengua y tecnologías de la información, (pàgs. 4–10). Novática: Revista de la Asociación de Técnicos de Informática.

McEnery, T.; Wilson, A. (1999). *Corpus Linguistics*. Edimburgo: Edimburgh University Press. *Resum*: Manual sobre les tècniques i fonaments bàsics de la investigació en lingüística de corpus textuals assistida per ordinador. Està concebut com un llibre de text amb exemples, exercicis i lectures.

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

Moré, J.; Climent, S.; Oliver, A.; Taulé, M. (2005). «Análisis de los fenómenos lingüísticos de los mensajes de correo electrónico en catalán desde la perspectiva de la traducción automática». A: *Procesamiento del Lenguaje Natural*, (35): pàgs. 45–50.

Moreno Fernández, F. (1994). «Status quaestionis: sociolingüística, estadística e informática». A: *Lingüística*, (6): pàgs. 95–154.

Ooi, V. (1998). *Computer Corpus Lexicography*. Edimburgo: Edimburgh University Press.

Pereira, F.; Warren, D. (1980). «Definite Clause Grammars for Language Analysis». A: *Artificial Intelligence*, (13): pàgs. 231–278.

Pérez Guerra, J. (1998). *Análisis computarizado de textos: una introducción a TACT*. Vigo: Universidade de Vigo (Servicio de Publicacións).

Pérez Hernández, C.; Moreno Ortiz, A.; Faber, P. (1999). «Lexicografia computacional y lexicografía de corpus». A: *Revista Española de Lingüística Aplicada*, volum 1: pàgs. 175–214. ISSN 0213-2028.

Pollard, C.; Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Standford: CSLI.

Ramallo, F. (1999). «Informática y sociolingüística cuantitativa». A: *Revista Española de Lingüística Aplicada*, volum 1: pàgs. 263–290. ISSN 0213-2028.

Reiter, E.; Dale, R. (1997). «Building Applied Natural Language Generation». A: *Natural Language Engineering*, (3): pàgs. 57–87.

Rosner, M.; Johnson, R. (eds.) (1996). *Computational Linguistics and Formal Semantics*. Cambridge: Cambridge University Press.

Ruiz Antón, J. C. (1996). *Modelos de análisis sintáctico en el procesamiento del lenguaje natural*, volum Lingüística e informática. Santiago de Compostela: Tórculo.

Shieber, S. (1986). *An introduction to unification-based approaches to grammar*, volum 4 de *Lecture Notes*. CSLI.

Shieber, S. (1998). *Separating linguistic analyses from linguistic theories*, volum Natural Language Parsing anf Linguistic Theories. Dordrecht: Kluwer.

Solias, T. (1996). *Gramática categorial: modelos y aplicaciones*. Madrid: Síntesis.

Sperberg-McQueen, M.; Burnard, L. (eds.) (1994). *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. ACL-ACH-ALLC. ISBN Chicago.

Strzalkowski, T. (ed.) (1999). *Natural Language Information Retrieval*. Kluwer. ISBN Dordrecht.

Tapias Merino, D. (1999). «Sistemas de reconocimiento de voz en las telecomunicaciones». A: *Revista de lingüística aplicada*, volum I: pàgs. 83–102.

Wanner, L. (ed.) (1996). *Lexical Fuinctions in Lexicography and Natural Language Processing*. Amsterdam: John Benjamins.

Way, A. (2001). «Translating with Examples». A: «Proceedings of MT Summint VIII Workshop on Example-Based Machine Translation», (pàgs. 66–80). Santiago de Compostela.

Whitelock, P.; Kilby, K. (1995). *Linguistics and Computational Techniques in Machine Translation System Design*. Londres: UCL Press. *Resum*: Anàlisi, comparació i avaluació de sis destacats sistemes de traducció automàtica. Els autors discuteixen diversos aspectes relacionats amb els problemes lingüístics de la traducció automàtica i presenten les tècniques bàsiques per al seu tractament informàtic.