Tecnologies del llenguatge

Anna Fernández
Glòria Vázquez

PID_00159190

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Introducció

Aquest és el primer mòdul de l'assignatura *Llengua catalana i tecnologies digitals*. Es tracta d'un mòdul introductori en què es plantejaran qüestions generals sobre l'àrea relacionada amb les tecnologies del llenguatge. En aquest mòdul trobareu informació sobre què designa aquest terme i quina relació té amb d'altres utilitzats dins l'àrea general en què s'emmarca aquest camp, els diferents camps de coneixement que hi intervenen i com interactuen, així com els processos, els models, els components i els recursos més habituals en els productes creats dins el camp de les tecnologies relacionades amb les llengües.

Objectius

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

1. Adquirir una visió preliminar de les tecnologies del llenguatge, el seu objecte d'estudi i els seus objectius.

2. Comprendre la interdisciplinarietat de la disciplina.

3. Delimitar l'àrea d'estudi de la disciplina en relació amb altres àrees afins.

4. Conèixer les aplicacions i els recursos més importants dins l'àmbit de les tecnologies del llenguatge.

5. Comprendre el concepte de *model del llenguatge* i quin tipus de modelització es requereix en els sistemes de processament del llenguatge natural.

1. Denominació de la disciplina

Per comprendre el que abasta l'àmbit de les tecnologies del llenguatge, cal que reflexionem sobre el paper del llenguatge en la nostra societat. Des de la filosofia i també la psicologia, molts estudiosos mantenen que el llenguatge és el mitjà a través del qual estructurem el nostre cervell i, per tant, el nostre pensament individual i col·lectiu. A la vegada, ens permet expressar allò que pensem, és a dir, ens permet comunicar-nos i, per tant, ser ésser socials.

D'altra banda, l'ésser humà també es caracteritza per la creativitat. L'evolució de la nostra societat incorpora avenços constants sense els quals el món actual no seria concebible. La impremta, l'electricitat, el telèfon, els ordinadors i Internet són només uns pocs dels molt invents dels humans que han canviat la nostra vida i el nostre món.

Actualment, una persona en qualsevol lloc del món i des de qualsevol ordinador connectat a la Xarxa pot accedir a una quantitat ingent d'informació, milions i milions de dades emmagatzemades en servidors de qualsevol altra part del món. Aquest fet ha revolucionat la concepció que tenim sobre les comunicacions i sobre la gestió de la informació. Mai no havia estat possible que les persones poguessin tenir accés a tanta informació i d'una manera tan ràpida. En tant que una part molt important d'aquesta informació és textual i està escrita en diversitat de llengües, la tecnologia ha anat evolucionant per tal de processar-la i facilitar-ne l'accés, així com per assistir l'humà en la creació i la difusió de continguts. Aquestes tecnologies relacionades amb les llengües ens permeten extreure el màxim valor possible de tots els avenços que tenim disponibles, de les eines al nostre abast per a la comunicació, de les noves maneres d'intercanviar i interactuar que tenen les persones del segle XXI.

L'àrea d'estudi denominada *tecnologies del llenguatge* és un camp relativament nou. Es tracta d'un àrea molt dinàmica, que es troba en evolució constant, en paral·lel a l'evolució de la tecnologia. És un camp en què treballen experts de diverses disciplines, tan diverses com poder ser la informàtica, la intel·ligència artificial, la matemàtica, la ciència cognitiva, la filosofia, la biologia i la lingüística. És precisament per aquest caràcter multidisciplinari i per la contínua aparició de noves tecnologies, així com també per interessos més enllà de la mateixa disciplina (Moreno Sandoval, 2008), que l'àrea d'estudi que ens ocupa ha estat denominada de diverses maneres al llarg dels anys.

Al llarg dels aproximadament seixanta anys d'existència de programes informàtics que tracten dades lingüístiques han aparegut diferents termes per designar aquesta disciplina. A continuació, n'esmentem els més comuns:

- *enginyeria lingüística*

- *indústries de la llengua*
- *informàtica (aplicada a la) lingüística*
- *lexicometria*
- *lingüística computacional*
- *lingüística informàtica*
- *lingüística quantitativa*
- *processament de dades lingüístiques*
- *processament del llenguatge natural*
- *tecnologies de la parla*
- *tecnologies del llenguatge (humà)*
- *tecnologies lingüístiques*

Alguns d'aquests termes s'han utilitzat, i encara s'utilitzen a vegades indistintament, però, a mesura que s'ha anat desenvolupant aquest camp d'estudi, els diferents termes s'han anat especialitzant per designar les diferents àrees d'investigació i d'aplicació o bé, en alguns casos, han deixat de ser utilitzats. Tanmateix, cal dir que les fronteres entre aquestes diferents àrees encara presenten zones borroses. Seguidament s'aniran perfilant els àmbits que abraça cadascuna. En realitat, els termes esmentats es poden agrupar de la manera següent:

- **Àrees d'estudi:** lingüística computacional, processament de dades lingüístiques, processament del llenguatge natural, informàtica (aplicada a la lingüística).

- **Tractament superficial de textos:** lexicometria, lingüística informàtica, lingüística quantitativa.

- **Aplicacions i programes comercials:** enginyeria lingüística, indústries de la llengua, tecnologies del llenguatge (humà), tecnologies lingüístiques.

- **Tractament de textos orals:** tecnologies de la parla.

En els apartats següents (del 2 al 5) anirem presentant cadascun d'aquests camps, que tracten de maneres diverses les relacions entre les llengües i les tecnologies. A l'apartat 6 ens estendrem amb la descripció dels processos, els models, els components i els recursos més habituals en el camp que ens ocupa. Finalment, presentem un apartat amb unes reflexions sobre els continguts tractats en aquest mòdul, seguit d'un petit recull de portals i adreces d'Internet que poden ser interessants per a aquelles persones que es vulguin endinsar en aquesta àrea.

2. Àrees d'estudi

En el camp de la lingüística, és a partir de la dècada dels seixanta que es comença a distingir entre lingüística teòrica i lingüística aplicada. Aquesta distinció no pretén reflectir una dissociació entre ambdues, ja que, com en el cas de les altres ciències, el vessant teòric i pràctic estan interrelacionats (Fernández, 1996).

Es podria dir que la lingüística aplicada estudia aquells aspectes de la teoria lingüística que poden donar resposta a unes necessitats socials (Payrató, 1997). En aquest camp, s'estableixen connexions entre la lingüística i altres ciències que estan implicades en la resolució d'aquests problemes. Des d'aquest punt de vista, la lingüística aplicada és un vessant de l'anomenada *lingüística externa* o *interseccional*.

Dins de la lingüística aplicada s'inclou tradicionalment l'ensenyament de llengües, la traducció, també l'anomenada *lingüística clínica* (en què conflueixen la fonètica, la neurolingüística i la psicolingüística) i, en menor o major mesura, la sociolingüística. A poc a poc, la lingüística aplicada va anar obrint les seves portes cap a la incorporació d'altres disciplines emergents en l'àmbit i que tenen en comú l'objectiu de resoldre problemes reals relacionats amb el llenguatge. Com a conseqüència, al llarg de la darrera dècada ja va començar a acceptar-se la incorporació del camp de la lingüística computacional i l'enginyeria lingüística. Prova d'això és, per exemple, que alguns manuals que es van publicar a finals del segle XX en l'àmbit de la lingüística aplicada recullen ja seccions dedicades a la lingüística computacional. Un exemple n'és l'obra coordinada per Fernández (1996).

Sociolingüística

Una de les aplicacions de la sociolingüística és, per exemple, el disseny i la implementació de projectes de planificació lingüística.

Inicialment, però, aquest camp va ser concebut amb certa autonomia respecte de la disciplina de la lingüística. Un dels motius va ser que va néixer en el si d'una altra àrea de coneixement, la informàtica. D'altra banda, un altre factor que va contribuir a concebre la lingüística computacional com a externa als estudis lingüístics és el fet que la informàtica no és una ciència humana ni social, com ho són la psicologia o la sociologia.

Així, doncs, les dues disciplines que fonamentalment interactuen en l'àmbit que ens ocupa són la **lingüística** i la **informàtica.** Aquestes, al seu torn, es nodreixen d'altres àrees de coneixement per assolir els objectius plantejats en aquest nou camp d'estudi. De fet, per poder interactuar dues disciplines tan diferents i poder assolir els fins comuns, cal que cada una s'adapti metodològicament i en els seus propis objectius a l'altra.

La lingüística, d'una banda, ha d'adoptar teories formals que permetin formalitzar els fets lingüístics, tot establint comportaments regulars i tot expressant-los amb un llenguatge inequívoc. D'una altra banda, la informàtica ha treballat tradicionalment amb llenguatges artificials i, per tant, si ha de processar el llenguatge natural, haurà d'adaptar-se per tal de tractar-ne adequadament les especificitats. Pensem que el llenguatge natural, a diferència dels llenguatges artificials, té ambigüitats i irregularitats, a les quals haurà de donar resposta.

L'interès per la disciplina va néixer a mitjan segle XX, moment en què la informàtica, que s'ocupa del processament automàtic d'informació, s'endinsa en el terreny de les llengües naturals, en tant que són codis utilitzats per transmetre informació. Es parteix de la creença que tant la ment humana com els ordinadors són processadors d'informació que poden manipular signes (és a dir, entitats amb un significat i un significant associats per convenció) i realitzar processos complexos (com inferir dades, prendre decisions, aprendre coneixement nou, etc). Per aquest motiu, i amb l'optimisme dels anys cinquanta davant els avenços al món de la informàtica, es va creure que era possible reproduir els processos cognitius de la ment de l'ésser humà en els ordinadors (Meya i Huber, 1986). Així va aparèixer aquest camp d'estudi.

Grishman (1986) defineix el terme *lingüística computacional*, traduït directament de l'anglès (*computational linguistics*), com "el estudio de los sistemas de computación utilizados para la comprensión y la generación de las lenguas naturales" (pàg. 15). D'altra banda, el *processament del llenguatge natural* és definit per Allen (1987) com "the basic techniques that are used in building computer modules of natural language production and comprehension" (pàg. 1).

Tal com es pot observar a partir d'aquestes definicions, la lingüística computacional i el processament del llenguatge natural són descrits de manera pràcticament idèntica, per la qual cosa es pot dir que poden usar-se ambdós termes indistintament. Amb el temps, però, l'ús dels termes *lingüística computacional* i *processament del llenguatge natural* s'han especialitzat segons l'àmbit des del qual es treballa. Així, dins l'àmbit de la lingüística s'utilitza preferentment –però no exclusivament– el terme *lingüística computacional*, ja que, tal com es reflecteix en l'estructura d'aquest sintagma, es posa èmfasi en el llenguatge vist des d'una perspectiva computacional. En informàtica, en canvi, se sol parlar de *processament del llenguatge natural*, ja que aquesta disciplina s'ocupa del

tractament (processament) de dades, que en aquest cas són de tipus lingüístic. A vegades, però de manera menys habitual, s'utilitza també la denominació *informàtica (aplicada a la) lingüística* per referir-se a aquest darrer àmbit.

El terme *processament del llenguatge* abraça, de fet, una àrea més àmplia. Aquest terme prové de l'àmbit de la psicolingüística, en què s'estudia quin tipus de fenòmens mentals intervenen en la comprensió i la producció lingüística. En el moment en què els informàtics pretenen que els ordinadors processin dades lingüístiques, és a dir, que les analitzin (comprensió) i les generin (producció), es planteja la creació de sistemes que emulin el processament humà del llenguatge. Com fins a aquest moment els llenguatges amb què treballen els informàtics eren artificials, es fa necessari afegir l'adjectiu *natural*, i per això el *terme processament del llenguatge natural*.

Cal dir, però, que, en el moment en què va néixer la lingüística computacional, la informàtica va prescindir de la teoria lingüística i es va alimentar únicament de tècniques dissenyades per al tractament dels llenguatges artificials –especialment els llenguatges de programació. Amb el pas del temps no solament es va establir connexió amb la lingüística teòrica, sinó que també es va veure que el processament del llenguatge natural implicava la utilització de processos intel·ligents i que una simple adaptació de les tècniques esmentades no conduïa a res. Així és com la lingüística computacional va entrar en contacte amb la *intel·ligència artificial*. Aquesta és una àrea de coneixement en què es desenvolupen sistemes informàtics que mostren una conducta intel·ligent i que, per tant, tracta de codificar en programes facultats cognitives, entre elles, la facultat lingüística. Cal dir que, de fet, el terme *lingüística computacional* com a tal va ser encunyat per David Hays, el qual l'entenia com una branca de la intel·ligència artificial.

Com ja s'ha comentat, en aquesta disciplina treballen conjuntament especialistes de cada un dels àmbits implicats. Els professionals d'una de les disciplines, des de la seva formació especialitzada, s'aproximen a l'altra disciplina per tenir-ne una versió àmplia sobre els objectius i la metodologia més bàsica. Aquest tipus d'aproximació és fonamental per establir un clima de cooperació i enteniment entre els diferents especialistes.

En general, els objectius de la disciplina s'aconsegueixen mitjançant dos tipus de processos:

a) L'aplicació del paradigma computacional a l'estudi científic del llenguatge humà.

b) El desenvolupament de sistemes informàtics per processar (comprendre i generar) els textos escrits o orals.

El primer d'aquests processos se situa en l'àmbit més lingüístic que computacional i el segon se situa en l'àmbit més computacional que lingüístic. Pel que fa al primer procés, el lingüista que treballa en l'àmbit de la lingüística computacional formalitza les dades lingüístiques amb llenguatges basats en formalismes gramaticals. En el segon procés, els informàtics tradueixen aquests formalismes a algoritmes i programes elaborats amb llenguatges propis de la disciplina de la informàtica. La separació de les dues àrees és crucial, ja que codificar directament el coneixement gramatical en un programa és molt costós i, a més, suposaria que els lingüistes fossin programadors.

Un dels primer llenguatges que els lingüistes van utilitzar per formalitzar i que es va usar en aquest camp va ser el Prolog, que prové de la lògica de predicats. D'altra banda, gràcies a l'aplicació de mètodes matemàtics a la lingüística, es va aconseguir un grau elevat de formalització de les llengües naturals des d'aquesta mateixa disciplina. Entre les aportacions més importants a la formalització del llenguatge cal destacar les aportades des del corrent generativista, que es nodreix en els seus inicis de la lingüística distribucional. Aquesta línia no va portar al desenvolupament d'aplicacions funcionals però sí que va obrir camí a altres teories lingüístiques, com les gramàtiques d'unificació, que han permès l'evolució d'aquesta àrea d'estudi pel que fa a la formalització.

En aquest sentit, les propostes de la lingüística teòrica són utilitzades en la lingüística computacional. De fet, la lingüística computacional i la lingüística teòrica comparteixen el mateix objectiu, la descripció i l'explicació dels processos lingüístics, encara que amb fins diferents. Quant a la lingüística teòrica, l'objectiu esmentat és el fi últim, ja que com qualsevol ciència, la lingüística vol arribar a caracteritzar el seu objecte d'estudi, en aquest cas, el llenguatge, i donar compte de les regles generals que són subjacents a les llengües i el comportament lingüístic. Pel que fa a la lingüística computacional, el que es pretén és descriure el funcionament lingüístic humà amb la finalitat d'intentar aconseguir amb mitjans informàtics els mateixos resultats, si pot ser, i idealment, emulant el comportament cognitiu humà. Per tal d'aconseguir aquest objectiu comú, des dels dos vessants es requereix dissenyar un *model* (arquetip) de com és el funcionament dels processos lingüístics. Si el model ha de reproduir tan bé com pugui les característiques de l'objecte i també tots els processos que hi actuen o bé ha de tenir un caire més operatiu, és una qüestió que estudiarem més endavant.

Vegeu també

Les funcions d'aquest model es tractaran al subapartat 6.1. "Els processos i els models".

Hem vist com la lingüística computacional es nodreix dels avenços en la lingüística teòrica per avançar en el terreny de les formalitzacions. Ara bé, la influència entre la lingüística teòrica i computacional és mútua, ja que els resultats obtinguts en aquesta última reverteixen també en la millora dels resultats de la primera.

En primer lloc, la lingüística computacional ha facilitat la utilització d'un *mètode empirista*. Després de dècades en què la lingüística teòrica s'havia centrat en l'ús d'un mètode únicament deductiu, basat, en general, en la introspecció, en l'actualitat, el lingüista pot accedir a grans quantitats de dades lingüístiques, els corpus en format electrònic, que contenen textos, tant orals com escrits, de varietats diverses (geogràfiques, socials o estilístiques).

Vegeu també

El mòdul "El processament de corpus" està dedicat precisament als corpus en format electrònic.

En segon lloc, la lingüística computacional pot ser utilitzada com a *banc de proves* per a la lingüística teòrica. Els models computacionals descriuen el comportament lingüístic per processar textos a partir dels formalismes que es creen en l'àmbit de la lingüística teòrica. La implementació i la utilització amb èxit d'aquests models teòrics per al processament de textos, sobretot si no pertanyen a àmbits concrets de coneixement, atorguen validesa a aquests models; i a l'inrevés, els resultats negatius de la implementació del model poden donar llum sobre aspectes susceptibles de millora en la teoria.

Finalment, no volem acabar aquest apartat en què ens hem endinsat en les especificitats del camp d'estudi relacionat amb les tecnologies lingüístiques sense advertir que el fet que un lingüista *usi* eines informàtiques relacionades amb el llenguatge no implica que *sigui* un lingüista computacional, ja que, encara que coneix l'ús adequat d'aquestes eines, no necessàriament comprèn els mecanismes, les tècniques i els mètodes subjacents. En canvi, una característica distintiva del lingüista computacional és que té coneixements i habilitats per entendre aquests mecanismes i també per dissenyar-los i desenvolupar-los.

3. Tractament superficial de textos

El terme *lingüística informàtica* sol utilitzar-se per referir-se als sistemes que no inclouen coneixement lingüístic i realitzen un tractament superficial del text (escrit). La *lexicometria* és un altre terme relacionat amb aquesta àrea, que bàsicament realitza una anàlisi quantitativa dels textos.

Aquest tipus d'anàlisi sembla que es remunta al segle XVIII. Com a exemples de l'ús d'aquesta tècnica s'esmenta el recompte que va fer un diari dels Estats Units dels elements bàsics del discurs antifederalista perfecte segons les vegades que apareixien els mots *llibertat de premsa, esclavitud* i *aristocràcia*, entre d'altres. No obstant això, no es considera que es pugui considerar l'anàlisi textual com a procediment científic fins a principis del segle XX. La incorporació de la informàtica a l'anàlisi textual és relativament recent i ha suposat, òbviament, un gran avenç quant a la possibilitat d'analitzar grans quantitats de material escrit.

En els programes lexicomètrics, els textos són considerats com a seqüències de caràcters, ja sigui de tipus lingüístic, numèric o de qualsevol altre tipus, independentment de la llengua que es tracti. L'enfocament de l'estudi dels textos és, doncs, quantitatiu. Per això, també s'utilitza el terme *lingüística quantitativa* com a sinònim de *lingüística informàtica*.

Amb aquest tipus d'eines també es poden obtenir dades estadístiques per a determinades seqüències de mots. Així, doncs, no es tracta d'un recompte simple de paraules sinó que es té en compte l'entorn en què aquestes s'usen per tal de conèixer-ne el comportament textual. Per mitjà d'aquests programes es poden obtenir, doncs, dades de la dimensió sintagmàtica de cada forma del text.

Es poden diferenciar dos tipus bàsics de procediments usats per aquests programes: els **documentals** i els **estadístics.** Els primers reorganitzen les unitats textuals, és a dir, les presenten d'una manera diferent del text lineal. El tipus de dades que s'obtenen mitjançant aquests processos d'anàlisi són:

- la llista de totes les paraules diferents que apareixen en el text
- el nombre de vegades que apareix una cadena en un corpus
- les paraules que comencen/acaben/contenen una determinada seqüència de caràcters
- les col·locacions i locucions
- els contextos en què apareixen els mots, tant per la dreta com per l'esquerra (concordances)

Els segons, els estadístics, realitzen diferents càlculs d'índexs de freqüència i comparacions entre diferents textos d'un corpus.

Procediments estadístics

Mitjançant els procediments estadístics es poden, per exemple, comparar les diferents parts d'un text segons el vocabulari usat i la freqüència d'ús d'aquest per tal d'establir similituds o diferències entre diferents emissors o entre diferents textos del mateix emissor i per tal d'establir quin vocabulari tendeix a ser molt o poc usat en un text o en determinades parts d'un text.

Els programes usats en la lingüística quantitativa són molt útils en àrees molt diverses, tant en l'àmbit dels estudis filològics, ja sigui literaris com estilístics, com també més pròpiament lingüístics, sobretot per al camp de la lexicografia, com veurem més endavant, però són eines que es poden usar en altres àmbits. Així, es poden aplicar mètodes lexicomètrics per analitzar dades del mateix autor recollides a través del temps i es poden posar en evidència les variacions que s'han produït. Aquesta tècnica s'ha usat no solament en el camp de la literatura, sinó també per analitzar textos polítics i per a l'estudi de processos psicològics, sempre que es disposi de dades textuals dels subjectes implicats recollits al llarg del temps, com diaris personals, cartes, entrevistes enregistrades, etc.

En realitat, la frontera entre la lingüística quantitativa –que fa una anàlisi superficial del text– i la lingüística computacional –que utilitza coneixement lingüístic en el processament de les dades– no és tan clara.

D'una banda, la lingüística computacional beneficia la lingüística quantitativa, ja que processos com l'anàlisi morfològica, sintàctica i semàntica dels textos milloren les possibilitats de cerques en els programes de la lingüística quantitativa. Això es pot observar amb més claredat en l'àmbit de la lingüística de corpus. Així, com més rica sigui la informació que s'anota en un corpus lingüísticament més ho seran les dades que poden aportar als estudis lingüístics que, al seu torn, estan basats en anàlisi de corpus.

Seguint aquesta idea, si els textos estan anotats amb informació morfosintàctica o d'algun altre tipus, es poden fer cerques complexes tot combinant les cerques de seqüències de caràcters amb etiquetes usades en l'anotació, com ara les categories morfosintàctiques. Aquest tipus de cerques proveeixen informació sobre la combinació d'elements que són crucials per determinar-ne el grau de "solidaritat lèxica" que presenten, tot calculant fins a quin punt hi ha una relació estreta entre aquests elements comparant els resultats obtinguts amb les dades de freqüència de cada un dels elements de manera aïllada. Per això es pot extreure, per exemple, el tipus de preposicions que acompanyen típicament determinats verbs per tal de deduir règims preposicionals, o bé la llista d'adjectius que típicament acompanyen un determinant nom, per tal de detectar col·locacions.

Solidaritat lèxica

Aquest terme va ser encunyat pel lingüista Coseriu per referir-se a l'associació entre mots diferents. Aquest fenomen ha estat anomenat de maneres diverses en el camp de la lingüística: *lexies complexes, coocurrències, coaparicions, combinacions, col·locacions, veïnatge*, entre d'altres. En el camp del processament del llenguatge natural també s'usa *termes multiparaula*.

D'altra banda, els resultats obtinguts a partir de l'anàlisi quantitativa dels textos també permeten millorar els sistemes del processament del llenguatge natural. Per exemple, permeten alimentar la informació associada a les entrades lèxiques dels diccionaris o lèxics que inclouen els diversos sistemes de processament, ja sigui amb informació relativa a les col·locacions, règims preposicionals o estructures oracionals. Aquesta informació pot ser molt útil, per exemple, en la traducció i correcció automàtiques de textos.

Vegeu també

Els sistemes de processament es tractaran al subapartat 6.4.

4. Aplicacions i programes comercials

Els termes *enginyeria lingüística, indústries de la llengua, tecnologies del llenguatge (humà)* i *tecnologies lingüístiques* van ser encunyats durant els anys noranta en el marc dels programes d'investigació de la Unió Europea per designar les aplicacions que permeten fer visibles a la societat els avenços aconseguits en l'àrea de la lingüística computacional i, per tant, aquelles aplicacions que tenen una finalitat pràctica i són susceptibles de ser comercialitzades.

Els usuaris finals d'aquests programes poden ser experts en llengües o no experts. Dins els **usuaris no experts en llengües**, inclouríem el públic general o sectors específics.

Quant al *públic en general*, qualsevol pot ser un usuari potencial d'un programa de correcció assistida o de traducció automàtica, ja sigui com un producte comercial adquirit o incorporat en altres recursos, com ara portals d'Internet, processadors de textos, gestors de missatgeria, etc.

Vegeu també

Als mòduls "Recursos d'ajut a l'edició" i "La traducció automàtica" podeu trobar més informació sobre els programes de correcció assistida i els programes de traducció automàtica, respectivament.

El públic no especialitzat també pot ser usuari dels productes de tecnologia de la parla, com ara de programes de dictat automàtic o bé d'interfícies de pregunta-resposta per telefonia. Actualment s'estan desenvolupant interfícies multimodals que permeten la comunicació amb un ordinador combinant la veu, la imatge, el vídeo i els gestos.

A més, hi ha moltes eines a la Xarxa que un usuari qualsevol pot utilitzar per a la cerca i la gestió d'informació. Aquesta àrea, la recuperació de continguts, té com a objectiu ajudar l'usuari en la cerca de documents complets o extractes d'aquests. Hi ha moltes aplicacions d'interès per a l'usuari final i una de les més conegudes per al públic en general és la dels cercadors d'informació a través d'Internet.

Vegeu també

Les eines per a la cerca i la gestió d'informació es tracten al mòdul "Cerca i recuperació d'informació".

Les eines de cerca d'informació són de gran utilitat també en àmbits específics, com el de la *documentació* i la *biblioteconomia*, àrees des de les quals s'han fet avenços molt importants en aquesta línia. També hi ha aplicacions amb relació a la recuperació d'informació dins del *món empresarial* per ajudar a la presa de decisions tot analitzant patrons de comportament i perfils de possibles clients, per exemple. Altres eines relacionades amb aquest camp són les interfícies de pregunta-resposta en *l'àmbit comercial* per a transaccions, reserves, etc.

Dins el camp de la recuperació d'informació, també hi ha eines de resum automàtic, que es poden usar des del processador de textos o bé des d'aplicacions més sofisticades via web. En aquestes aplicacions se sol treballar identificant els segments més rellevants del text.

En general, l'objectiu últim en aquesta àrea és fer arribar la informació d'interès per a l'usuari i, si pot ser, de manera que pugui ser gestionada i utilitzada àgilment. Així, en qualsevol àmbit en què es tractin grans volums de dades estructurades la recuperació d'informació serà d'interès.

Pel que fa als **usuaris experts en llengües**, cal destacar sobretot els traductors professionals, els lexicògrafs i els terminòlegs, així com, en major o menor mesura, els professors de llengües. Aquests professionals són usuaris potencials de lèxics digitalitzats, bases de dades terminològiques, eines d'extracció de terminologia (eines que proposen candidats a *termes* a partir de l'anàlisi estadística, normalment, de corpus especialitzats) i programes de concordances. Aquest tipus de programes, doncs, no tenen interès per al públic general, però es consideren que pertanyen a l'àmbit de l'enginyeria lingüística perquè estan dirigides a professionals que són susceptibles de consumir aquests productes per millorar les tasques que duen a terme en la seva vida laboral.

Els traductors professionals, sobretot de textos especialitzats, són un perfil de professional de la llengua que clarament s'ha bolcat a l'ús de tecnologies lingüístiques, concretament de les eines relacionades amb la traducció assistida, que inclouen memòries de traducció (és a dir, fragments ja traduïts que s'usen per fer una primera traducció parcial automàtica del text) i bases terminològiques. Aquests professionals són també usuaris en major o menor mesura dels programes de traducció automàtica disponibles a la Xarxa o bé també en les versions comercials. Gairebé tothom ha utilitzat alguna vegada algun sistema de traducció en línia dels que s'ofereixen a Internet. Tots sabem que no són perfectes però la incorporació de les memòries de traducció en aquests sistemes darrerament ha millorat en alguns casos els resultats obtinguts. Per a un usuari no expert moltes vegades la funcionalitat bàsica d'aquests programes es limita al fet que ens puguem fer una idea del contingut d'un text i, en el cas dels traductors professionals, els pot servir com a punt de partida.

Vegeu també

Recordeu que la traducció assistida es tracta al mòdul "La traducció automàtica".

5. Tecnologies de la parla

Tradicionalment, els investigadors de l'àmbit del processament del llenguatge natural han treballat únicament amb textos escrits. Això obeeix principalment a raons històriques, ja que el processament del llenguatge natural es va iniciar des de l'òptica de la informàtica, mentre que el del tractament de la parla va començar en el marc de l'enginyeria de telecomunicacions.

Així, doncs, el camp relacionat amb el processament de senyals sonors del llenguatge sol ser tractat de manera autònoma i rep el nom de *tecnologies de la parla*.

Fins fa uns anys, el processament de la parla i el text s'estudiaven independentment. Avui dia, encara que s'accepta que l'àmbit de la lingüística computacional pot incloure també el tractament de textos orals i hi ha congressos en què els investigadors aporten resultats des dels dos vessants, aquesta no és una convenció totalment consensuada.

En aquesta assignatura no es tractarà dels aspectes relacionats amb la tecnologia de la parla. Aquesta disciplina es pot considerar en ella mateixa un camp independent, en tant que la problemàtica que s'afegeix pel que fa al tractament del so, tant pel que fa al reconeixement de cadenes fòniques com pel que fa a la producció sonora (síntesi de veu), pertany ben bé a una altra àrea, més lligada amb la física, l'enginyeria de telecomunicacions i, des del punt de vista lingüístic, a la fonètica acústica.

Un cop reconeguda una cadena fònica, caldrà associar-la amb un text escrit i, a partir d'aquí, caldrà usar les tècniques del processament escrit. El procediment contrari consisteix a produir un text sonor a partir d'una cadena de caràcters del codi escrit. Els problemes a què han de fer front aquests sistemes tenen a veure amb aspectes com la discriminació dels sons propis del llenguatge d'altres tipus de sons o com el reconeixement de diferents varietats dialectals (geogràfiques i socials), així com la producció de cadenes fòniques amb una entonació adequada, en el cas dels sistemes de síntesi de veu.

En el camp que ens ocupa, és rellevant el vessant aplicat de la investigació que s'hi ha dut a terme i, en aquest sentit, els productes creats en aquest àmbit constitueixen una part del que hem anomenat l'*àrea de l'enginyeria lingüística*, que, com ja hem avançat, inclou els programes amb aplicacions comercials.

Una aplicació prototípica d'aquest àmbit són els programes de dictat i els sintetitzadors de veu pensats sobretot per a discapacitats, així com també en el camp de l'ensenyament de llengües pel que fa a la correcció fonètica.

Una altra aplicació prototípica dins el món de les tecnologies de la parla són els sistemes de diàleg que s'utilitzen en diverses empreses i institucions per atendre les trucades telefòniques. Dins aquests sistemes, la complexitat pot ser més o menys elevada, i això dependrà del grau de coneixement que incorpora el sistema i del model que s'ha usat per construir l'aplicació. Així, hi ha sistemes que consisteixen únicament a donar opcions a l'oient sense que aquest participi amb la veu, o si ho fa, sempre a partir d'un nombre tancat d'opcions. D'altra banda, hi ha sistemes que tracten d'emular una conversa amb l'usuari i que, inicialment, no li presenten limitacions en la construcció de les emissions que produeix. En aquests casos, es requereixen sistemes que incorporen algun tipus d'interpretació semàntica. Els problemes propis de la gestió del diàleg que també han de resoldre aquests sistemes tenen a veure amb el tractament de les inferències i de la pragmàtica. Així, doncs, són de natura semblant, encara que amb especificitats pròpies de la llengua oral, als sistemes de diàleg per a textos escrits.

Vegeu també

Els sistemes de diàleg per a textos escrits es tracten al mòdul "Cerca i recuperació d'informació".

Cal dir que els resultats en l'àmbit de les tecnologies de la parla són encoratjadors i els avenços en aquesta àrea són molt importants. Darrerament s'han aconseguit ja algunes fites en aquest camp, com és la traducció de la parla espontània, que és, de fet, la forma que tenim els humans de comunicar-nos per excel·lència. Així, avui dia podem dir que ja hi ha aplicacions en què es tradueixen les produccions orals sense estar emmarcades en un context específic, és a dir, sense estar restringides a un context comunicatiu.

6. Estructura dels sistemes de processament de les llengües

En aquest apartat descriurem de manera genèrica com es construeixen els sistemes que processen dades lingüístiques, més concretament, quins processos han de realitzar i a partir de quins models es configuren i de quins mòduls i components estan formats. A més, s'esmentaran els recursos lingüístics i els programes informàtics més rellevants que solen utilitzar-se en les diferents aplicacions existents en el camp de les tecnologies del llenguatge i se'n descriuran les característiques principals. Finalment, per a cada tipus d'aplicació s'explicaran quins de tots aquests elements que conformen l'estructura d'un sistema de processament lingüístic estan en joc.

6.1. Els processos i els models

Les aplicacions creades en el camp de les tecnologies del llenguatge poden realitzar bàsicament dos tipus de processos, que es corresponen amb els dos processos bàsics del comportament lingüístic humà: l'**anàlisi** i la **generació d'informació.**

L'**anàlisi** consisteix a transformar les dades lingüístiques en elements de coneixement representats de manera que puguin ser tractats per un sistema intel·ligent. Es tracta de convertir una sèrie de símbols (el llenguatge escrit) o senyals (el llenguatge parlat) en afirmacions sobre objectes estructurats segons un sistema de representació del coneixement processable.

La **generació** és el procés invers de l'anàlisi, ja que ara els objectes estructurats (representació) es converteixen en elements lingüístics.

Exemple

En un sistema de traducció automàtica, la fase d'anàlisi consisteix a transformar el text d'una llengua en una representació semàntica intermèdia entre les dues llengües. En el llenguatge humà aquesta fase és la de comprensió d'un missatge.

En canvi, en un sistema de traducció automàtica, la fase de generació consisteix a transformar la representació intermèdia en el text de la llengua de destinació. En el llenguatge humà la fase corresponent és la de la producció de missatges.

Hi ha aplicacions que usen els dos procediments, com hem vist amb la traducció automàtica, i també succeeix amb els sistemes de diàleg home-màquina, mentre que d'altres només integren un dels dos processos i/o no sempre de manera completa. Així un generador d'informes meteorològics és una eina que és capaç de *produir* textos en llenguatge natural a partir de símbols obtinguts d'un satèl·lit. D'altra banda, una eina de correcció de textos sol *analitzar*

parcialment la llengua escrita, sense arribar a comprendre els missatges. Al llarg d'aquesta assignatura revisarem diverses aplicacions i com hi tenen lloc aquests processos d'anàlisi i generació de dades, però sobretot ens centrarem en els primers i en menor mesura en els segons.

Com ja hem esmentat, l'objectiu de la disciplina que ens ocupa és dissenyar un **model** que sigui capaç de generar i/o analitzar textos i es pugui implementar informàticament. La naturalesa del model pot ser diversa, en funció de fonamentacions de base procedimental i filosòfica.

A l'inici del desenvolupament de la lingüística com a ciència, a principis del segle XX, es va considerar que aquest model bàsicament havia de donar compte d'aspectes estructurals (morfologia i sintaxi, sobretot). Algunes d'aquestes teories incloïen la semàntica i d'altres la van incorporar ja ben avançat el segle XX. No serà fins als anys seixanta que es va començar a estendre la necessitat que el model fos complert, i que inclogués també els aspectes funcionals. Així, és en aquest moment quan van començar a prendre importància els estudis de pragmàtica, que tenen com a focus d'interès l'ús del llenguatge, que té en compte el context comunicatiu en què tenen lloc els intercanvis lingüístics. Mentre que la formalització dels nivells foneticofonològic, morfològic i sintàctic podem dir que està plenament o gairebé assolida en el segle XXI, en la formalització de la semàntica i la pragmàtica encara hi ha terreny per treballar.

En la lingüística computacional, el model ideal ha de ser complet, és a dir, ha d'incloure tant els aspectes estructurals com els funcionals. Una aplicació que consisteix en un sistema de diàleg entre un humà i un ordinador a través d'un canal oral i obert a qualsevol domini (és a dir, no dissenyat per ser utilitzat només dins una temàtica i una situació comunicativa específiques) constituiria un exemple de sistema que requeriria un model d'aquest tipus. Aquest model idealment hauria d'incloure la formalització de tots els nivells lingüístics (des de la fonètica i la fonologia fins a la pragmàtica) i fins i tot més enllà de la lingüística, tot incorporant el coneixement del món. Així, si una màquina ha de mantenir, per exemple, una conversa amb un humà, ha de tenir el coneixement adequat en tots els nivells lingüístics per tal de poder actuar correctament. Per tant, davant d'una pregunta com "Pots dir-me quantes ciutats de més de 10.000 habitants hi ha a Catalunya?", esperem que el sistema no ens respongui si realment ens ho pot dir o no, sinó que volem que ens doni el nombre de ciutats que estem sol·licitant.

Un dels dilemes de la intel·ligència artificial és si ha de modelitzar formes humanes de tractament de problemes d'acord amb el que proposa la ciència cognitiva i, en el cas del llenguatge, la psicolingüística, o ha de limitar-se a obtenir resultats similars als que els éssers humans obtindrien si fessin front als mateixos problemes, tot explorant els seus propis mètodes de processament.

A la pràctica, molts dels sistemes creats donen resultats satisfactoris sense necessàriament aplicar models cognitius, sinó que incorporen processos d'altre tipus dissenyats *ad hoc* per donar compte dels problemes específics que ha de resoldre l'eina que s'està dissenyant. Així, el més habitual és que el model utilitzat i el seu abast estigui en relació directa amb l'aplicació que es vol construir i les necessitats que se'n deriven.

Podríem dir que en lingüística computacional hi ha dos tipus de models més comunament utilitzats: d'una banda, **els models basats en el coneixement** i, de l'altra, **els probabilístics** (o estadístics o estocàstics[1]).

(1)En el camp de l'estadística, un procés estocàstic és un procés aleatori.

Durant els anys setanta, es van fer aportacions importants des de l'àrea de la intel·ligència artificial al camp del processament del llenguatge natural per al desenvolupament del primer tipus de models especialment en el camp de la representació semàntica. En aquest camp va destacar l'aportació de Schank (1975) amb els anomenats *scripts* (guions), que es poden definir com a esquemes que inclouen seqüències d'accions predeterminades i estereotipades que defineixen una situació quotidiana.

La lògica difusa

Hi ha un altre model menys estès en el camp que ens ocupa basat en les xarxes neuronals i que utilitza els procediments de la lògica difusa. S'ha usat sobretot en l'àmbit de la recuperació de la informació, la mineria de dades i l'aprenentatge automàtic.

Quant a la sintaxi, en aquest tipus de models s'utilitzen sobretot gramàtiques d'estructura de la frase, per a la construcció de les quals es parteix del fet que les llengües són sistemes formals i que es poden definir amb un conjunt de *símbols* que descriuen les peces lèxiques i un conjunt de regles que expliciten les possibles combinacions dels símbols entre ells (estructures sintagmàtiques). Per això, se'n parla genèricament com a *models simbòlics*, que en realitat es poden considerar com un subtipus dels models basats en el coneixement, ja que per a determinades tasques, com el tractament d'inferències, es requereixen altres procediments que van més enllà de les regles de combinatòria de símbols.

Aquest tipus de formalització va ser especialment desenvolupada des del corrent generativista i l'exponent n'és Chomsky. Els conceptes subjacents són el del *poder generatiu* de les llengües i el de *competència*, ja que el que es pretén és donar compte de les possibles estructures *gramaticals* que es poden generar en una determinada llengua. En realitat, aquests models responen, doncs, a la filosofia del corrent racionalista, ja que pretenen *explicar* el comportament lingüístic.

En la dècada dels vuitanta aquestes gramàtiques d'estructura sintagmàtica es van anar sofisticant i van aparèixer formalismes gramaticals més elaborats que fan ús de trets i restriccions. Tot aquest tipus de formalismes han tingut un paper especialment rellevant en els processos de comprensió (anàlisi) del llenguatge.

Però durant els anys vuitanta també neixen els *models probabilístics* i tindran un protagonisme especial en la dècada dels noranta, sobretot en el camp de les tecnologies de la parla, però no exclusivament. Durant aquests anys es desfocalitza l'interès per les teories i hi ha un gir cap a les aplicacions. Això implica que el que interessa són els resultats i, en conseqüència, es comença a qüestionar la necessitat d'emular els processos mentals, ja que el que es pretén és que l'ordinador executi les tasques sense pretendre que s'apropi a la ment humana. Això fa que apareguin aquests nous models com a alternativa als basats en el coneixement, que havien estat els més usats. Aquests models probabilístics es fonamenten en tècniques d'*aprenentatge* a través de corpus (*machine learning*), és a dir, el que es pretén és inferir coneixement a partir de dades. Així, doncs, des del punt de vista del corrent filosòfic, estan ancorats en l'empirisme. En el vessant lingüístic, això implica que el que és pretén és *descriure* el comportament lingüístic, és a dir, *l'actuació* i, per tant, no se centren en la gramaticalitat de les oracions sinó en com són les llengües des de l'*ús*.

Lectures complementàries

Per a un visió dels models estadístics aplicats a la lingüística vegeu:

E. Charniak (1993). *Statistical Language Learning*. Cambridge: MIT.

E. Charniak (2000). "A maximum-entropy-inspired parser". A: *Proceedings of NAACL-2000* (pàg. 132-139).

Els mètodes quantitatius utilitzats en aquests sistemes es basen en la idea que es poden inferir les estructures d'una llengua a partir de la cerca en un corpus de regularitats basades en l'estadística. Donada una seqüència d'elements, es busquen patrons de coaparició en el que s'anomenen els *n-grames*, que són seqüències de *n* elements. Les regularitats es poden basar en la copresència de determinades *unitats lèxiques*. Aquesta tècnica és la que es fa anar, per exemple, quan es pretén desambiguar el sentit dels mots polisèmics d'un text (desambiguació de sentits del mot), ja que a l'hora de prendre la decisió es tenen en compte amb quins mots coapareix en el discurs la forma analitzada. En cas que es pretengui realitzar una anàlisi morfosintàctica i que es disposi de corpus anotats, l'extracció de regularitats es basa en la proximitat o coaparició de determinades *categories morfosintàctiques lèxiques o sintagmàtiques*.

Desambiguació

Aquest camp, denominat en anglès *word sense disambiguation,* ha estat un dels àmbits amb més desenvolupament dintre del processament del llenguatge natural en les dues darreres dècades.

En la taula 1, adaptada de Church i Mercer (1993), es presenta un resum de les principals diferències entre ambdós tipus de models. Tant un tipus de model com l'altre no són perfectes i presenten alguns problemes. Entre els principals inconvenients de l'ús de models basats en el coneixement en sistemes de processament automàtic de textos destaquem la menor *robustesa*, en comparació dels sistemes que es basen en models estadístics. Diem que un sistema és robust quan és capaç de resoldre un important nombre de casuístiques diverses. Aquest problema es deriva de la creativitat del llenguatge i es dóna sobretot quan el domini no és especialitzat. Tot i que les llengües es poden descriure i els lingüistes poden donar compte de les regles subjacents a les oracions, es donen casos de construccions sintàctiques poc previsibles, però possibles, i que sovint no són descrites en les gramàtiques per la manca de representativitat. Tot i que això suposa un problema per als models basats en el coneixement, té solució, ja que consisteix a augmentar el nombre de regles previstes. Encara que aquest afegitó disminueix l'elegància del model, actua en pro de la resolució de problemes reals. D'altra banda, el maneig de gramàtiques amb un nombre de dades important és cada vegada menys problemàtic des del punt

Lectura complementària

K. Church; R. Mercer (1993). "Introduction to the Special Issue on Computational Linguistics Using Large Corpora". *Computational Linguistics* (vol. 19, núm. 1, pàg. 1-24).

de vista de l'eficiència informàtica, llevat de les aplicacions que funcionin de manera interactiva en temps real, com en els productes relacionats amb les tecnologies de la parla.

Taula 1. Comparació entre els models basats en el coneixement i els probabilístics

	Models basats en el coneixement	**Models probabilístics**
Corrent filosòfic	Racionalisme	Empirisme
Model	Competència	Actuació
Objectius	Gramaticalitat, explicació	Descripció
Aplicacions	Comprensió	Tecnologies de la parla, cerca d'informació
Elements	Estructura sintagmàtica	*n*-grames

El problema es complica a l'hora de resoldre els problemes d'ambigüitat, que caracteritzen les llengües naturals i que es donen a diferents nivells lingüístics. Si pensem en una oració com

El conferenciant va parlar als professors de tecnologia

veurem que a aquesta oració se li podria atorgar una doble anàlisi a partir del subconjunt de regles següent d'una gramàtica del català:

1) SN → DET N

2) SN → DET N SP

3) SN → N

4) SV → V SP

5) SV → V SP SP

6) SP → PREP SN

Si apliquéssim la regla 1 per al subjecte i les regles 4, 6, 2, 6 i 3 per al predicat, el resultat seria:

$[El\ conferenciant]_{SN}\ [va\ parlar\ [a[ls\ professors]_{SN}\ [de\ [tecnologia]_{SN}]_{SP}]_{SP}]_{SV}$

En canvi, si en lloc de la regla 4 utilitzéssim la 5 i en lloc de la regla 2 utilitzéssim la 1, el resultat seria:

$[El\ conferenciant]_{SN}\ [va\ parlar\ [a[ls\ professors]_{SN}]_{SP}\ [de\ [tecnologia]_{SN}]_{SP}]_{SV}$

En el primer cas, la interpretació que obtindríem és que els professors són experts en tecnologia i, en el segon cas, la tecnologia seria la temàtica del discurs del conferenciant i el públic serien docents sense una especialitat especificada. És a dir, en el segon cas el predicat conté dos complements, mentre que en el primer només en conté un.

L'humà és capaç de desambiguar l'estructura sintàctica subjacent d'aquesta frase i, per tant, el significat de l'oració, tenint en compte el context semantico-pragmàtic. Ara bé, els models simbòlics troben limitacions en aquesta tasca de desambiguació. Recordem que aquests models es van crear per a la resolució dels aspectes lingüístics relacionats amb la morfologia i la sintaxi, però en canvi no estan pensats per a la implementació de regles semàntiques ni pragmàtiques. Si la resolució de l'ambigüitat es pot fer afegint context a la regla, és a dir, complicant la sintaxi de la mateixa regla, és factible, però si requereix necessàriament incorporar coneixement d'altre tipus, aquests models no sempre aconsegueixen donar sortida als problemes lingüístics.

Una de les limitacions més importants en aquest sentit és com formalitzar el coneixement del món, és a dir, com donar compte de com funciona el món en què vivim. Aquest tipus d'informació és ingent i es tracta més aviat d'informació de caire enciclopèdic. Habitualment s'utilitzen ontologies per reflectir la conceptualització que es troba subjacent a l'expressió lingüística però determinat tipus de coneixements sobre com es relacionen els objectes i les entitats són difícils de reflectir a través de l'ús de formalismes i jerarquies de conceptes.

Els models probabilístics, com hem avançat, són més robustos i, per tant, són capaços d'adjudicar una anàlisi, la més plausible segons els seus càlculs, a les diferents frases dels textos, la qual cosa no implica que la solució que aportin sigui l'encertada en tots els casos. Per tant, les limitacions dels models probabilístics giren entorn d'altres problemes. Així, aquests models es fonamenten en les dades reals que s'exploren a partir de l'anàlisi de textos, ja que la idea és inferir el comportament lingüístic usant-los com a font. Si els textos estan en brut (sense anotació de cap tipus), el tipus de regularitats que se'n poden extreure, com hem vist, està limitat bàsicament a la semàntica lèxica i al tractament dels termes multiparaula. Per tant, per poder avançar de manera completa en aquest àmbit, es requereixen textos enriquits amb informació lingüística de diversa índole (morfosintàctica, sintàctica, semàntica o pragmàtica).

D'altra banda, els resultats que obtindrem en aplicar models probabilístics estan limitats també pel tipus de fenòmens reflectits en els textos que s'analitzen. És evident que com més grans siguin els textos en què es basen els algoritmes per crear les gramàtiques d'anàlisi, més bons resultats s'obtindran, però ja hem esmentat que l'anotació d'informació lingüística no és fàcil ni ràpida i, a més, un corpus, per molt gran que sigui, mai no es pot considerar com una representació completa d'una llengua. Això vol dir que és molt factible que, en analitzar textos nous, les gramàtiques creades a partir de la casuística

d'uns determinats textos, trobin esculls amb noves estructures. En aquest cas, la limitació no es troba en la capacitat de descripció del lingüista, com en els models simbòlics, sinó en la varietat que es troba en els textos usats com a punt de partida per crear l'eina que servirà d'anàlisi per a textos futurs.

Finalment, tots els fenòmens lingüístics que no es presentin en la frase de manera contínua presenten una limitació important per als models de què estem tractant. Alguns exemples de discontinuïtat en les llengües són algunes oracions interrogatives en què l'element interrogatiu està expressat de manera separada del constituent al qual pertany. Els models simbòlics, encara que també estan basats prototípicament en la presència contigua de símbols, presenten estratègies per superar aquest tipus de fenòmens, com en el cas de les gramàtiques categorials, en què s'opta per l'ús d'operadors específics per tractar els casos de discontinuïtat.

Discontinuïtat gramatical

Un exemple de discontinuïtat gramatical és el cas de *quina història* a *Quina història que va explicar el Pere sap la Maria?* Aquest constituent apareix separat del verb del qual depèn, *saber*, amb una seqüència de caràcters entre ells i, fins i tot, amb un altre verb interposat (*explicar*).

En tot cas, tant els models basats en el coneixement com els probabilístics es troben amb un problema comú, que és la resolució d'ambigüitats que pràcticament solament es poden resoldre a través del que hem anomenat *coneixement del món*. Així, en la frase *Volaré de Roma a Tòquio en dues hores*, l'única interpretació possible per a un humà adult estàndard seria que l'avió s'enlairarà d'aquí a dues hores, ja que no és possible la interpretació segons la qual l'avió trigarà dues hores per fer el trajecte esmentat. En aquest cas l'ambigüitat no és de tipus sintàctic, ja que l'estructura sintagmàtica és sempre la mateixa, sinó semàntic, atès que en un cas el complement temporal es refereix al moment inicial de l'acció de volar i en l'altre cas es refereix a la durada total del vol. Per als models basats en el coneixement, la tria d'una interpretació o una altra es veu limitada pel tipus de coneixement que solen incorporar aquests sistemes. D'altra banda, com hem vist, els models probabilístics poden prioritzar una interpretació, la que determinin com a més freqüent, però en tot cas hi ha un marge d'error.

En la darrera dècada, arran de les limitacions que s'han anat evidenciant per als models probabilístics, s'ha apostat per un ressorgiment dels models simbòlics, tot combinant-los amb els primers i tot aprofitant de cadascun els avantatges que presenten en pro d'una millor eficiència en la resolució de tasques complexes. Sembla, però, que encara no hi ha una metodologia establerta que determini quina combinació de mètodes és la més apropiada per a cada aplicació. S'han dut a terme experiments diversos que combinen ambdues tècniques, com ara l'ús de mètodes estadístics per ampliar el coneixement dels formalismes gramaticals basats en models simbòlics. La idea és barrejar diferents tipus de coneixement i sembla que aquesta és l'aposta de futur.

6.2. Els components i els mòduls

A continuació, ens aturarem en els models que es basen en coneixement lingüístic i tractarem de descriure'ls de manera global, centrant-nos en els processos d'anàlisi de textos i pensant en aplicacions ideals que requereixen el

processament de dades lingüístiques a diversos nivells, excepte el foneticofonològic, que va més enllà dels objectius que ens hem marcat en l'assignatura, com ja s'ha argumentat.

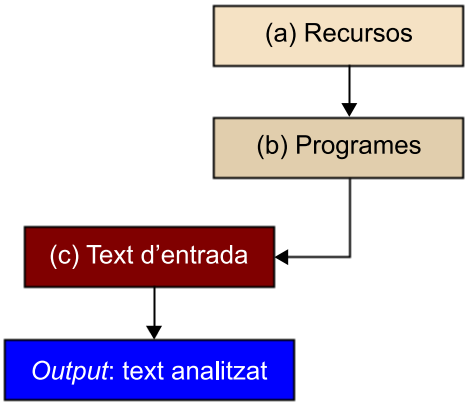
En un sistema de processament del llenguatge natural basat en aquest tipus de model hi ha tres components bàsics, com es pot veure a la figura 1:

- **Recursos (a):** són dades generals de la llengua degudament formalitzades per tal que es puguin computar. Es tracta habitualment de gramàtiques i lexicons en què es codifica la informació necessària sobre el llenguatge per tal d'obtenir els resultats esperats un cop s'hagi efectuat el procés (aquestes gramàtiques i lexicons s'anomenen *recursos del sistema).*

- **Programes (b):** són eines informàtiques que utilitzen les dades lingüístiques (a) sobre els textos que es volen analitzar (c).

- **Text d'entrada (c):** són l'objecte del procés i el que es pretén analitzar amb un major o menor grau de profunditat, segons els objectius de l'aplicació.

Figura 1. Components bàsics d'un sistema de processament de la llengua basat en un model amb coneixement lingüístic



El tipus de dades de (a) o el tipus de programes de (b), la manera en què aquests components interactuen, els mòduls de què es compon el sistema i el tipus de textos que tracten (c) varia molt d'un sistema a un altre i també depèn del tipus d'aplicació. L'objectiu últim és aconseguir que l'ordinador sigui capaç de "comprendre" textos escrits, és a dir, que n'aporti una anàlisi i, a vegades, una interpretació (*output*).

Donada la complexitat de la tasca que es pretén abraçar, els sistemes es dissenyen modularment. Així, de la mateixa manera que en lingüística es distingeixen diferents nivells en l'anàlisi del llenguatge (fonètic, fonològic, morfològic, sintàctic, semàntic i pragmàtic), els sistemes que processen els textos parteixen igualment d'una **concepció modular**, de manera que cada mòdul s'encarrega d'un determinat tipus d'unitats (sons, fonemes, morfemes, paraules, frases, discurs) i un programa gestor decideix quin actua en cada moment.

L'ús d'un sistema modular té avantatges des del punt de vista informàtic, ja que el processament del llenguatge és un problema complex que posa en joc unitats de naturalesa diversa. Vegem quins són els avantatges principals que s'han destacat d'aquest tipus d'estructura:

- Permet mantenir cada mòdul de manera independent i, per tant, es poden aplicar programes específics per al tractament de les unitats de cada nivell: les unitats i el procés que s'aplica per a l'anàlisi morfològica són diferents dels que s'utilitzen, per exemple, en la interpretació semàntica.

- Permet mantenir la independència interna de les dades, de manera que es poden realitzar modificacions en un mòdul sense que això hagi d'afectar la resta; això és important perquè un sistema sigui flexible i ampliable.

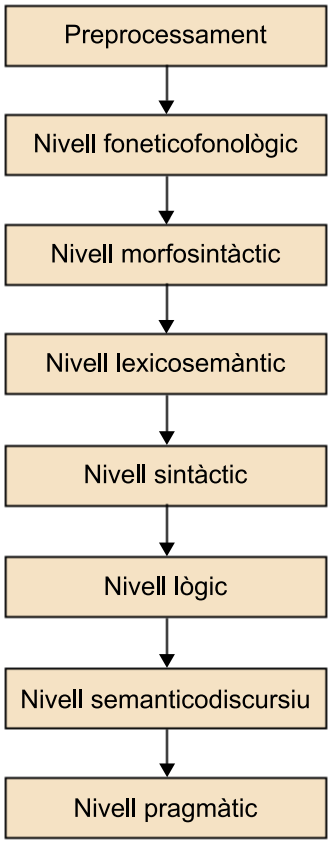
- Permet el desenvolupament simultani de cadascun dels mòduls durant la construcció del sistema i l'especialització dels diferents grups de treball.

- Permet la integració efectiva dels diferents nivells mitjançant la definició del que cada component espera a l'entrada i del que produeix com a sortida; cal, per tant, construir programes que garanteixin la consistència interna de les dades.

S'acostumen a presentar els mòduls de manera estratificada, des dels més pròxims a la realització superficial fins als que estan relacionats amb les capacitats cognitives de més alt nivell. Cal recordar que els nivells fonètic i fonològic es tracten en els sistemes de síntesi i reconeixement de la veu. Aquests nivells no són previstos quan l'aplicació se cenyeix al text escrit, com en el cas que ens ocupa. En el cas dels sistemes de síntesi de veu, la funció d'aquest nivell serà la de convertir una cadena de fonemes en sons i, a partir d'aquí, materialitzar-los físicament. En el cas del reconeixement de la veu, caldrà partir de la cadena sonora per tal d'arribar a representar els fonemes que hi són subjacents i a la cadena ortogràfica corresponent.

En la figura 2 s'esquematitza com podria ser l'arquitectura en mòduls d'un sistema que analitza i interpreta textos escrits.

Figura 2. Estructura modular d'un sistema d'anàlisi i interpretació de textos



La delimitació de l'abast de cada mòdul pot variar. Així es pot considerar que l'anàlisi morfosintàctica es pot fer en un nivell i que la desambiguació correspon a un altre mòdul. En tot cas, l'important és la construcció gradual de la interpretació.

1) Preprocessament

En aquesta fase es procedeix a la identificació de títols, dates, noms propis, unitats multiparaula, etc. Cada un d'aquest tipus d'elements quedarà classificat de manera que el sistema actuarà amb aquestes peces d'informació segons s'estipuli en cada nivell. Per exemple, es pot decidir no processar els noms propis i, en canvi, sí considerar les unitats multiparaula.

2) Nivell morfosintàctic

En aquest nivell es tracta la segmentació interna de les unitats mínimes del text en lexemes, morfemes flexius i, si cal, derivatius, i s'assigna a cadascun dels elements les seves possibles interpretacions morfosintàctiques. En una primera fase, en cas que un mot pugui tenir més d'una possible anàlisi si es considera aïllat del context, se n'indiquen totes. En una segona fase, s'utilitzen regles per desfer les ambigüitats, tenint en compte el context.

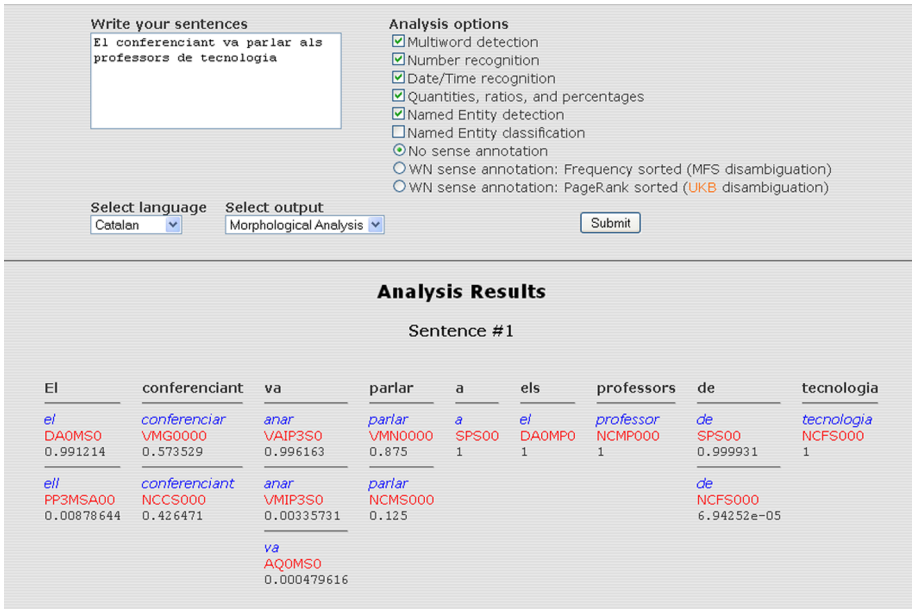
Com veiem en la figura 3, alguns dels mots de la frase que hem pres com a exemple anteriorment (*El conferenciant va parlar als professors de tecnologia*) presenten ambigüitat morfosintàctica. Bàsicament, s'hi poden identificar tres ambigüitats principals[2]: en primer lloc, el mot *el* pot ser un determinant (DA0MS0) o un pronom (PP3MSA00); en segon lloc, el mot *conferenciant* pot ser un gerundi (VMG0000) o un nom (NCCS000), i, en tercer lloc, el mot *va* pot correspondre's amb diferents formes verbals del verb *anar* (VAIP3S0, auxiliar, i VMIP3S0, present d'indicatiu) i amb l'adjectiu masculí singular (AQ0MS0). Per a cada anàlisi, es presenta a més un índex numèric que indica la freqüència d'ús de cada possible anàlisi. En els casos en què no hi ha ambigüitat el valor és 1.

[2]Com es pot observar en la imatge, la forma *de* s'associa també a dues possibles formes: la preposició, clarament més freqüent, i el nom metalingüístic de la preposició. Per a la forma *parlar* també es tenen en compte dues possibilitats: la del verb en infinitiu, també molt més freqüent, i la de la nominalització de la forma verbal esmentada.

Analitzador FreeLing

Aquest és el resultat de l'analitzador FreeLing (Atserias i altres, 2006), desenvolupat per investigadors de la Universitat Politècnica de Catalunya. En l'adreça següent es pot consultar el paquet d'eines d'anàlisi morfològica i sintàctica que posa a disposició aquest grup i que s'ha usat per extreure les anàlisis presentades tant en aquesta figura com en les que es presenten més endavant (figures 4, 5 i 6): http://garraf.epsevg.upc.es/freeling/demo.php.

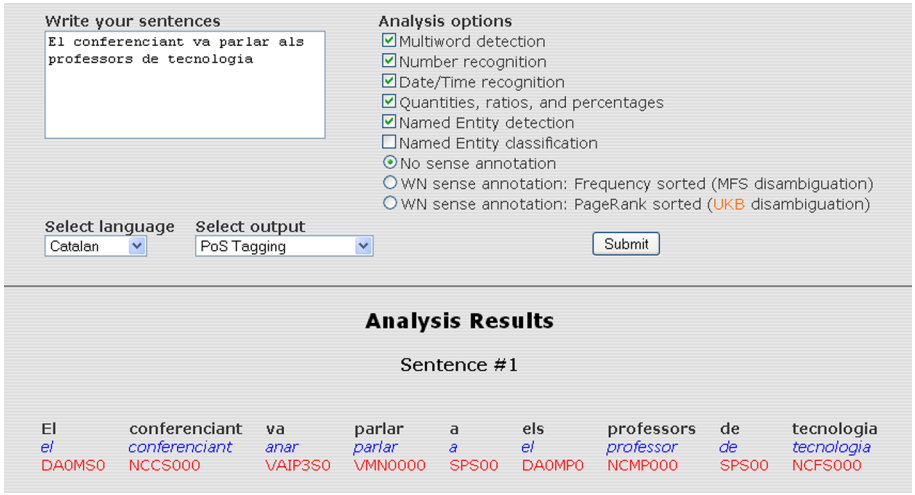
Figura 3. Resultat de l'anàlisi morfològica de la frase *El conferenciant va parlar als professors de tecnologia*



En una segona fase, s'utilitza una altra eina per desambiguar en aquells casos en què l'analitzador ha proposat més d'una possible interpretació. El sistema realitza la tria en funció del context en què es troba cada peça lèxica i la freqüència de coaparició. Així, com es pot veure en la figura 4, durant aquest procés de desambiguació[3] s'ha optat per les solucions següents: en primer lloc, per a la forma *va* el sistema ha triat l'auxiliar, ja que darrere hi ha un verb en infinitiu; en segon lloc, per a la forma *conferenciant*, el sistema ha triat la forma nominal, ja que un gerundi no apareix al costat de la forma *el*, independentment que aquesta sigui determinant o pronom, i, finalment, per a la forma *el*, s'ha escollit el determinant, ja que al costat d'aquest mot hi ha un nom.

[3]En la figura el procés esmentat és anomenat *POS tagging*, ja que, d'una banda, les eines de desambiguació en anglès s'anomenen *taggers*, en tant que trien l'etiqueta (*tag*) més adient per caracteritzar el mot, i d'una altra banda, aquesta etiqueta es refereix a la categoria morfosintàctica, que en anglès rep el nom habitualment de *part-of-speech* (POS).

Figura 4. Resultat de la desambiguació morfològica de la frase *El conferenciant va parlar als professors de tecnologia*



3) Nivell lexicosemàntic

En aquest nivell s'assigna a cada unitat lèxica identificada tota la informació que conté el lema corresponent. Si el lèxic té la informació es pot afegir una descomposició del significat. En tot cas, per als noms, es pot afegir el tipus semàntic i, per als verbs, es poden aportar dades sobre l'estructura argumental, com ara el nombre d'arguments, els tipus de sintagmes i els rols semàntics dels participants.

Així, davant una frase com la que estem analitzant, *El conferenciant va parlar als professors de tecnologia*, la informació sobre l'estructura argumental del verb *parlar* extreta del lèxic seria que aquest predicat pot tenir com a complements un SP introduït per la preposició *a* (destinatari) i un SP introduït per la preposició *de*[4] (tema). Tot i que aquesta informació és molt útil per a l'anàlisi sintàctica i la interpretació semàntica de l'oració, no és habitual que els lèxics usats en les aplicacions comercials incloguin aquest tipus de dades, ja que el fet de descriure les especificitats lèxiques per a les llengües requereix una inversió important (de temps i diners).

[4]Aquest verb pot presentar també altres configuracions argumentals.

4) Nivell sintàctic

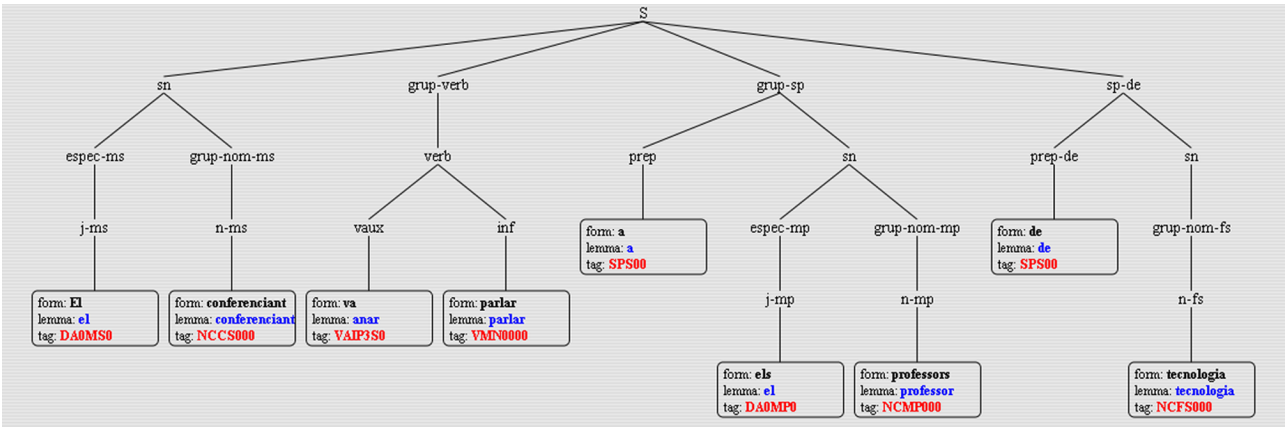
L'anàlisi sintàctica es realitza partint de l'*output* generat en el nivell morfosintàctic i bàsicament pot efectuar-se de dues maneres. En el primer cas, es reflecteixen les interrelacions dels mots en el si de les frases identificant-ne els constituents. L'anàlisi sintàctica resultant en aquests casos pot ser parcial o total. En l'anàlisi parcial s'identifiquen els sintagmes però no s'arriba a establir sempre la relació jeràrquica entre aquests. Així, per a una frase com la ja esmentada (*El conferenciant va parlar als professors de tecnologia*), el resultat seria:

$$[\text{El conferenciant}]_{SN} \text{ va parlar } [\text{a}[\text{ls professors}]_{SN}]_{SP} [\text{de } [\text{tecnologia}]_{SN}]_{SP}$$

Com es pot veure, s'agrupen els mots en sintagmes bàsics, però no s'estableixen relacions de dependències entre aquests. En la figura 5 es presenta en forma de diagrama arbori aquesta anàlisi.

Si l'anàlisi fos total s'obtindrien una o dues representacions sintàctiques, en funció de si es resolgués o no l'ambigüitat estructural relacionada amb l'SP *de tecnologia*, respectivament. En l'anàlisi total s'estableix quins sintagmes configuren l'SV i l'oració.

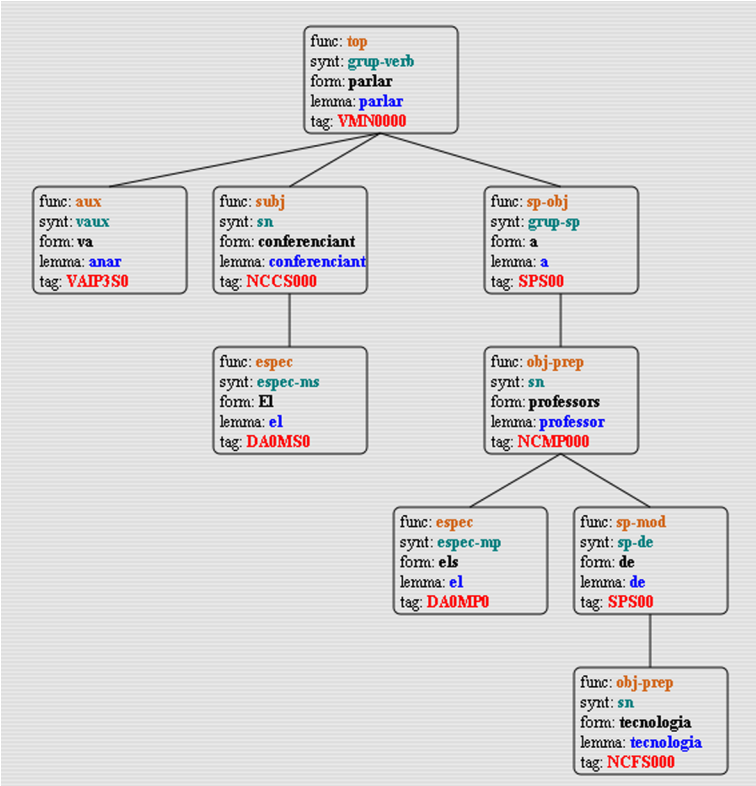
Figura 5. Resultat de l'anàlisi sintàctica parcial basada en estructures sintagmàtiques de la frase *El conferenciant va parlar als professors de tecnologia*



En aquest tipus d'anàlisi també es poden incorporar les funcions sintàctiques si el sistema és prou potent per identificar-les (Marcus i altres, 1994), tot i que per a determinades llengües, com, per exemple, la catalana, continua essent una tasca difícil de resoldre, ja que és una llengua d'ordre més lliure que, per exemple, l'anglès.

A més de l'anàlisi basada en sintagmes, també hi ha eines que apliquen anàlisis basades en el model de dependents i regents de Tesnière. El resultat s'exemplifica en la figura 6.

Figura 6. Resultat de l'anàlisi sintàctica basada en dependències de la frase *El conferenciant va parlar als professors de tecnologia*



5) Nivell lògic

En aquest nivell s'obté una interpretació de la frase sense tenir en compte el context extralingüístic ni el context lingüístic més enllà de la mateixa oració. Es parteix del resultat obtingut en l'anàlisi sintàctica. Per simplificar l'explicació, ens centrarem només en la representació lògica que obtindríem per a la frase que estem usant com a exemple, tot partint d'una de les possibles anàlisis sintàctiques que es podrien obtenir.

Així, de manera molt sistemàtica, podríem dir que el que s'expressa en la frase de l'exemple és que hi ha un acte de parla en el qual intervenen tres participants, que tenen les funcions semàntiques d'agent, destinatari (o beneficiari) i tema (informació que vindria codificada en l'entrada del verb *parlar* del component lèxic). En el llenguatge de la lògica dels predicats aquesta informació s'expressaria amb una representació en què el verb és el nucli de la fórmula, ja que expressa la semàntica de l'acte en si, i els participants esmentats són els arguments d'aquest predicat. Per al verb *parlar*, partiríem d'una fórmula lògica com la següent, proporcionada a partir de la informació del nivell lèxic:

(parlar [X,Y,Z])

on:

- X és un SN agent
- Y és un SP (preposició *a*) destinatari

- Z és un SP (preposició *de*) tema

Les lletres X, Y i Z són variables que s'instancien amb els sintagmes de la frase que s'està analitzant i que s'han identificat en la fase de l'anàlisi sintàctica. Com a resultat, obtindríem:

$$(\text{parlar } [\text{conferenciant}_{\text{agent}}, \text{professors}_{\text{destinatari}}, \text{tecnologia}_{\text{tema}}])$$

A més, en aquesta fórmula caldria afegir també informació sobre les coordenades espaciotemporals en què ha tingut lloc l'acte, si és que aquesta informació està expressada en l'oració, així com altre tipus d'informació que s'hi pugui trobar relativa a altres circumstàncies en què ha tingut lloc l'acció, com ara la forma en què s'ha dut a terme l'acte de parlar, per posar un exemple.

6) Nivell semanticodiscursiu

En aquest nivell s'utilitza la informació del context lingüístic per completar, si és necessari, la identificació d'alguns arguments. És en aquesta fase en la qual s'haurien d'identificar, si s'escau, els referents d'elements com els subjectes nuls i les anàfores (pronoms relatius, reflexius, personals, etc). La cerca de candidats, per motius obvis, es prioritza sempre en el context anterior a la frase. D'altra banda, una altra qüestió que cal tenir en compte és quin tipus de categoria sintagmàtica cal identificar en el discurs com a candidat a referent. Normalment, es tracta de noms, com en el cas del subjecte elíptic i molts pronoms, però també podrien ser adjectius.

Per a la identificació dels subjectes no expressats, en llengües com les romàniques una altra informació que s'usa com a punt de partida per tractar de resoldre la tasca és la referent a la persona i el nombre del verb de l'oració i a partir d'aquí s'extreuen possibles referents de les oracions situades anteriorment en el discurs.

Quant a la identificació dels referents, les tècniques són diferents en funció de si els pronoms són relatius o d'un altre tipus. En el primer cas, el referent es localitza sovint en la posició immediatament anterior a l'anàfora o molt a prop. En el segon cas, és a dir, per a la resta de pronoms, el referent pot trobar-se en un context més llunyà.

7) Nivell pragmàtic

En aquest nivell es connecten els elements identificats en les anàlisis anteriors amb els elements del món definit en l'aplicació, ja siguin propis d'un domini restringit o bé d'abast més general. A partir d'aquí i amb l'ús de programes que permetin inferir coneixement, el sistema pot arribar a deduir informació que no està present lingüísticament ni en la frase analitzada ni en el discurs. Com ja s'ha esmentat, la formalització del funcionament del món és una tasca

difícil i l'ús d'ontologies no sempre permet arribar a inferir el coneixement que els humans tenim sobre el món que ens envolta i que en moltes ocasions és cabdal per comprendre el significat d'una oració.

6.3. Els recursos i els programes

Per a l'execució de totes aquestes fases es poden usar recursos i programes de diferents tipus (com es mostra a la figura 1). Diferenciem entre recursos i programes en tant que els primers (*a*, en la figura 1) són estàtics, ja que inclouen dades, mentre que els segons (*b*, en la figura 1) són dinàmics, ja que usen aquestes dades per generar una anàlisi dels textos.

A continuació, esmentem els **recursos** més utilitzats en el camp de les tecnologies de la llengua:

- lexicó o diccionari,
- *tesaurus* i ontologia,
- gramàtica.

Un **lexicó o diccionari** és una llista de paraules, que pot ser la llista de paraules d'una llengua o bé les pertanyents a un domini específic de coneixement.

A part de les paraules, un lexicó computacional pot tenir associada informació de diversa índole. Aquesta informació serà determinada pel tipus d'aplicació per a la qual s'ha dissenyat aquest mòdul. El tipus d'informació que més comunament se sol introduir acompanyant les formes dels mots és la que pertany al nivell morfosintàctic, ja que la categoria i la informació sobre gènere, nombre i persona, segons escaigui, són molt útils en diverses tasques a diferents nivells del procés.

Els diccionaris o lexicons són a la base de qualsevol sistema informàtic que tingui com a objectiu processar el llenguatge, ja que la seva funció és la de guiar qualsevol procés que es basi en el reconeixement de les paraules, ja sigui anàlisi morfològica, sintàctica o semàntica, traducció, cerca d'informació, etc. Com més informació i més diverses siguin les dades explicitades en el lexicó, més robust i més possibilitats d'aplicació tindrà el sistema.

D'altra banda, els ***tesaurus*** i les **ontologies** no tracten els mots aïlladament sinó que estableixen relacions, sobretot d'hiperonímia i hiponímia, entre ells.

Inicialment la diferència entre ambdós recursos s'establia en la naturalesa de la unitat que es prenia com a elemental, que en els *tesaurus* era lingüística i en les ontologies era conceptual. Ara bé, actualment, la diferència entre els *tesaurus* i les ontologies no sempre és clara i els dos recursos tendeixen a confluir.

Els *tesaurus* han estat utilitzats sobretot en el camp de la documentació com a eina per obtenir el vocabulari per descriure els documents que es volen organitzar de cara a indexar-los i recuperar-los posteriorment. En canvi, les ontologies han estat utilitzades en el camp de les tecnologies del llenguatge sobretot en l'àrea de la desambiguació de sentits. En aquest àmbit, les més habituals són WordNet (WN) (Fellbaum, 1998; Miller, 2009) i MicroKosmos (Beale i altres, 1995). Les ontologies també s'usen en aplicacions que requereixen comprensió del text, com ara les interfícies home-màquina, sobretot quan es tracten dominis restringits. En aquests casos es creen ontologies específiques per a l'àmbit a partir de les quals es poden realitzar inferències i altres processos intel·ligents relacionats amb l'anàlisi semàntica dels textos.

Lectures complementàries

S. Beale; S. Nirenburg; K. Mahesh (1995). "Semantic Analysis in the Mikrokosmos Machine Translation Project". A: *Proceedings of the Second Symposium on Natural Language Processing* (SNLP-95) (pàg. 297-307). Bangkok: Kaser Sart University.

C. Fellbaum (ed.) (1998). *WordNet. An Electronic Lexical Database.* Cambridge: MIT.

Miller, G. A. (2009). "WordNet - About Us". *WordNet* [en línia]. Princeton University.

Finalment, les **gramàtiques computacionals** són gramàtiques formals que pretenen donar compte de les possibilitats de combinació dels diferents elements lèxics de les llengües per formar sintagmes i dels sintagmes per formar oracions.

Cal donar compte de la coordinació i de la subordinació, que són fenòmens que caracteritzen totes les llengües naturals, així com de possibles reduplicacions d'elements, discontinuïtats, etc. A més, cal que les gramàtiques continguin mecanismes per controlar la concordança dintre dels sintagmes, quan escau, així com la concordança entre subjecte i verb. Per fer-ho, se solen utilitzar estructures de trets i mecanismes d'unificació que donen compte de manera molt eficient d'aquest tipus de fenomen. Una tendència instaurada a finals del segle XX en la formalització lingüística ha consistit a augmentar el poder dels lèxics i dotar-los d'informació de tipus gramatical pel que fa al comportament verbal. La idea és que és el component lèxic el que guia el procés d'anàlisi i les gramàtiques estan subordinades a la informació que s'obté d'aquell recurs. En relació amb aquest tipus de procediment es va encunyar el terme *gramàtica lèxica*.

Els programes informàtics usats en els sistemes de processament de textos (*b*, en la figura 1) poden ser independents de la llengua i usen els recursos esmentats (lèxics i gramàtiques, bàsicament), que sí són específics per a cada llengua, per aportar una anàlisi dels textos.

A continuació, esmentem alguns dels **programes** més comuns en el camp de les tecnologies del llenguatge:

- analitzador morfològic i lematitzador,
- etiquetador/desambiguador morfosintàctic o *tagger*,
- analitzador sintàctic o *parser.*

Els **analitzadors morfològics** (vegeu la figura 3) tenen com a objectiu aportar les característiques morfosintàctiques de les unitats d'anàlisi del text (mots o segments multiparaula). A més, si són també **lematitzadors**, associen el mot al lema (masculí singular per als adjectius, infinitiu per als verbs, etc.). Aquesta informació és bàsica per a les etapes posteriors del processament, sobretot el nivell lexicosemàntic i sintàctic.

Quant al nivell lexicosemàntic, la consignació del lema com a part de l'anotació del text permet heretar informació directament d'aquesta unitat, que és informació compartida per totes les formes que hi estan relacionades (per exemple, totes les formes de la conjugació d'un mateix verb comparteixen la mateixa estructura argumental). Aquest fet permet simplificar els lèxics, ja que permet economitzar a l'hora de descriure el comportament de les diferents unitats que el componen. Quant al nivell sintàctic, és imprescindible conèixer les propietats morfosintàctiques dels mots per tal de poder generar una estructura de la frase.

En general, aquest tipus de programes han presentat molt bons resultats en l'àmbit del processament del llenguatge natural, ja que, dintre de les tasques que s'hi realitzen, el grau de complexitat en aquests casos és baix.

Normalment, els processos d'anàlisi morfològica consten de diferents fases. En primer lloc, es reconeixen físicament els mots (o cadenes de caràcters) que constitueixen l'*input* del procés d'anàlisi. Un cop segmentat el fragment que es pretén analitzar, s'ha de comprovar que els segments reconeguts són una forma de la llengua. Arribats en aquest punt els analitzadors morfològics poden funcionar de dues maneres. La primera manera és simple i consisteix a assignar la informació associada a cada forma lingüística des del lèxic (propietats morfosintàctiques i lema). La segona opció consisteix a realitzar processos de descomposició dels mots en temps real, tot partint d'un diccionari d'arrels,

d'un diccionari d'afixos i d'un mòdul de regles de combinació entre els uns i els altres. En aquest darrer cas, un cop s'estableix quina és l'arrel de la forma, s'associa al seu lema.

La funció d'aquest tipus de programes és aportar totes les possibles anàlisis morfosintàctiques dels mots, per tant, l'*output* que s'obté pot evidenciar casos d'ambigüitat, com ja hem vist.

Per tal de desambiguar les formes s'han creat els programes anomenats ***taggers*** (**desambiguadors o etiquetadors morfosintàctics**). Normalment, un *tagger* decideix quina etiqueta i quin lema assigna a una forma en funció de regles de desambiguació basades en el context (vegeu la figura 4). Aquestes regles poden ser de coneixement lingüístic o bé purament estadístic. El resultat en ambdós casos sol ser de molt bona qualitat i, per tant, el marge d'error és petit.

Els **analitzadors sintàctics** o ***parsers*** parteixen de l'*output* dels *taggers* i construeixen estructures de la frase en què els diferents mots es presenten relacionats entre ells jeràrquicament utilitzant les dades que confereix la gramàtica formal del sistema. Els tipus de relacions que es poden establir, com hem vist en les figures 5 i 6, poden ser diversos: d'estructura sintagmàtica o de dependència bàsicament.

A més, com hem esmentat també, alguns *parsers* són capaços d'associar funcions sintàctiques als elements de la frase.

Un dels problemes més difícils de resoldre per als *parsers* és l'ambigüitat estructural de les frases en llenguatge natural. Sovint aquest problema queda sense resoldre i s'arriba només a anàlisis parcials (*shallow parsing*). Així, per a llengües que presenten restriccions d'ordre importants entre els elements de la frase és més fàcil la construcció d'eines que permeten una anàlisi total. Per al castellà i el català, en canvi, se solen usar analitzadors parcials que realitzen l'agrupació bàsica en sintagmes. En aquest sentit, l'establiment complet de les dependències entre els elements d'aquestes llengües és un camp que encara no està del tot resolt.

6.4. Grau de complexitat de les aplicacions

Dins el camp de les tecnologies del llenguatge, es dissenyen i desenvolupen sistemes comercials que utilitzen recursos i tècniques de l'enginyeria lingüística que poden ser més o menys desenvolupats i poden tenir graus de complexitat i sofisticació molt diversos.

Per exemple, els **correctors automàtics** són una de les aplicacions més senzilles dintre el camp de l'enginyeria lingüística, ja que els objectius que es pretenen aconseguir amb aquest tipus de programes no són gaire ambiciosos. Així, la fita no és arribar a una anàlisi total del text suposadament incorrecte i la generació posterior del text correcte, sinó més aviat es parteix de la idea que l'eina *ajudarà* l'usuari a fer un text *millor*. Són, doncs, programes que assisteixen un usuari redactant textos tot alertant-lo de possibles problemes en l'escrit que produeix i, en la mesura del possible, proposant solucions.

El sistema pot tenir només un diccionari de formes de la llengua (una simple llista sense informació associada de cap tipus) i un programa que detecta les formes lèxiques del text que no formen part d'aquest lèxic tot comparant les seqüències de caràcters que escriu el redactor amb la llista de formes possibles per a l'idioma que s'està tractant. Alguns correctors detecten també alguns errors gramaticals bàsics, com la manca de concordança entre el determinant i el nom, però ho solen fer a partir del reconeixement de patrons gramaticals mínims i no solen incorporar ni gramàtiques ni analitzadors de cap mena. En aquests casos, els diccionaris que incorporen els correctors s'han d'alimentar amb informació morfosintàctica d'algun tipus que és reutilitzada en els patrons gramaticals.

En el cas de la **traducció**, les tasques que inicialment s'han de realitzar són d'un nivell de complexitat més elevat, ja que cal proposar un text de destinació i, per tant, es realitzen processos d'anàlisi per després poder generar. Tradicionalment, els sistemes de traducció automàtica es classifiquen segons el grau de coneixement lingüístic que incorporen. Previsiblement, els més sofisticats, que inicialment estan pensats per a llengües estructuralment molt diferents, presentarien una arquitectura per mòduls de processament semblant a la que hem descrit i, per tant, hi intervindrien diversos recursos, com ara lèxics amb informació sintacticosemàntica associada, analitzadors i generadors morfològics i sintàctics, a més d'alguna ontologia semàntica que permetés configurar estructures conceptuals de la llengua origen a partir de les quals es generarien els textos en la llengua de destinació. Es coneixen molt poques aplicacions que incorporin un processament tan complex. Actualment es tendeix a aprofitar les memòries de traducció com a pas previ de la traducció automàtica i també es barregen els mètodes probabilístics amb els basats en el coneixement lingüístic amb l'objectiu d'obtenir el millor rendiment amb un esforç menor.

Dins el camp del **processament de corpus**, ens trobem de nou amb diferents graus de complexitat. Recordem que un corpus és una recopilació gran d'instàncies de la llengua. Els corpus poden recollir dades orals o dades escrites, per tant, podem parlar de *corpus orals* i de *corpus escrits*, però en aquesta assignatura ens centrarem en el darrer tipus.

D'altra banda, els corpus es poden anotar amb informació o poden no anotarse (textos en brut). En el segon cas, evidentment, la construcció dels corpus és molt més simple i, per tant, el volum dels corpus pot ser molt més gran. En

el primer cas, el nivell de complexitat por variar. Per anotar un corpus amb informació lingüística podem utilitzar eines informàtiques que ens facilitin la tasca, ja que l'anotació manual és molt costosa. En el món del processament de corpus i, com a conseqüència de les polítiques establertes durant els anys noranta a la Unió Europea pel que fa a la creació de recursos lingüístics, s'usen estàndards per a l'anotació amb l'objectiu últim que les eines creades puguin compartir-se i explotar-se amb més facilitat. Aquests estàndards han estat molt estesos pel que fa al nivell textual dels documents, tot diferenciant entre les diferents parts d'aquest (títol, subtítol, paràgraf, etc.). Aquest és un pas molt útil per dur a terme el processament pròpiament lingüístic. En el camp de la morfosintaxi hi ha acords presos pel que fa a l'anotació, però en la resta de nivells lingüístics hi ha propostes diverses que no estan generalitzades.

EAGLES

El grup denominat Expert Advisory Group on Language Engineering Standards (EAGLES) es va crear a iniciativa de la Comissió Europea. L'objectiu d'aquest grup és la provisió d'estàndards per a la creació de recursos lèxics i corpus a gran escala i, en general, per als llenguatges de marcatge i altres eines per al tractament textual. El consorci NERC també va proposar uns estàndards per al marcatge morfosintàctic (TEI).

El grau d'automatització del procés d'anotació pot variar i depèn del tipus d'anotació que es vulgui dur a terme. El que cal tenir en compte és que és interessant processar els corpus de manera automàtica sempre que es pugui, sense que això impliqui un detriment de la qualitat de la informació lingüística, ja que l'interès principal que tenen els corpus, tant per a l'àrea de lingüística computacional com per a l'àrea de lingüística teòrica, és el d'aportar grans volums de dades per als estudis lingüístics o per a l'extracció d'informació estadística.

Quant a l'anotació manual, cal dir, però, que no està comprovat que sigui necessàriament de més bona qualitat. Així, encara que es consensuïn criteris per anotar determinats fenòmens lingüístics que es caracteritzen per certa complexitat, no està garantit que els anotadors actuïn sempre homogèniament, ja sigui per errades humanes o bé perquè també és possible que entrin en joc diferents interpretacions dels criteris consensuats segons els casos. En el cas de l'anotació automàtica, el criteri sempre és aplicat de la mateixa manera.

Aquells nivells d'anotació dels corpus que es poden automatitzar gairebé per complet serien l'anotació d'informació morfològica, en primer lloc, i, d'informació sintàctica, en segon lloc. Per a l'anotació morfosintàctica s'usen lèxics, analitzadors morfològics i *taggers*. Per a l'anotació sintàctica, es parteix de l'*output* generat en l'anàlisi morfosintàctica i s'usen els *parsers*.

Pel que fa al camp de la **cerca i gestió de la informació**, les tècniques més comunament usades en aquest àmbit pertanyen als models probabilístics i es basen en la coaparició dels mots de la cerca en un text, així com en la mesura de la densitat d'aquests mots en el document. Entre del recursos lingüístics utilitzats, cal esmentar sobretot les ontologies, per exemple, en les aplicacions creades per classificar documents segons l'àmbit temàtic al qual per-

tanyen. Dins del món dels cercadors, alguns usen tècniques mixtes que combinen la utilització de llenguatge natural i les tècniques estadístiques pròpies de la intel·ligència artificial. Entre el tipus de recursos lingüístics usats, a més de les ontologies, podem citar els analitzadors morfològics, ja que ambdós recursos es poden utilitzar per augmentar els possibles mots de la cerca i així millorar previsiblement els resultats obtinguts, ja sigui amb l'afegit de mots semànticament o morfològicament relacionats amb les paraules de la cerca triades per l'usuari.

7. Conclusions

En aquest mòdul hem vist que l'objectiu principal de la lingüística computacional és dissenyar com s'han de desenvolupar els sistemes i les aplicacions que tracten amb dades lingüístiques. Les tecnologies de la llengua són aquelles aplicacions que ens envolten i que tenen com a objectiu general que els humans es puguin comunicar i treballar de manera més eficient.

En els darrers anys s'ha observat una certa tendència envers la utilització de tècniques fonamentades en l'estadística i la probabilitat. Tot i que no s'ha demostrat que aquests models siguin perfectes, sí que han permès avançar en algunes àrees com són l'adquisició automàtica de grans volums de dades i la construcció d'eines d'anàlisi del llenguatge a partir de les dades recopilades. Sembla que la combinació entre aquests mètodes i els basats en el coneixement lingüístic poden ser el camí més encertat a l'hora d'afrontar els problemes que planteja el processament de les llengües amb vista a fer front als processos de comprensió i generació de manera *global*, tot tenint en compte tant els factors més estrictament pertanyents a la lingüística estructural com els relacionats amb la semàntica i, sobretot, la pragmàtica i el coneixement del món. En general, però, encara que aquest continua sent el marc de treball en el camp que ens ocupa, a poc a poc s'ha anat abandonant la idea d'arribar a un grau de comprensió total en el processament automatitzat de dades lingüístiques perquè s'ha pres consciència de les grans dificultats que això comporta.

Volem fer notar que, com en altres àmbits, es detecta un cert distanciament entre la comunitat científica i el sector empresarial en aquest camp. Així, d'una banda, els científics fan aportacions molt importants des del punt de vista teòric amb contribucions sobre els models de partida. D'una altra banda, molts equips investigadors desenvolupen recursos que, previsiblement, han de ser incorporats en productes finals. La creació d'aquests recursos es veu supeditada a l'existència del finançament, per la qual cosa, és recurrent que la cobertura de moltes d'aquestes eines no sigui completa. A més, en el sector empresarial, que és on es desenvolupen majorment les aplicacions, aquestes solen ser dissenyades amb objectius menys ambiciosos des del punt de vista de l'elegància del model però més ambiciosos quant a la resolució de problemes reals relacionats amb el processament del llenguatge.

Per acabar, podem dir que com a tendència general en l'àrea de les tecnologies del llenguatge, s'aposta per la presència cada vegada més palesa de més varietat de llengües. També, cada vegada més, es requereix que s'encavalquin les tecnologies de la parla i del text, ja que avui dia és habitual transmetre el coneixement mitjançant documents multimèdia, que incorporen so i text escrit, a més d'imatge.

Adreces d'Internet

A continuació, presentem una llista de llocs web en què es pot trobar informació sobre alguns dels temes que hem revisat en aquest mòdul i també recursos de tecnologies de la llengua.

1) Associacions i institucions

- **Association for Computational Linguistics (ACL):** http://www.aclweb.org. En aquest lloc web hi ha informació sobre els objectius d'aquesta associació, els congressos que organitza i la revista que publica.

- **Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN):** http://www.sepln.org/. Aquesta associació és la representació dels professionals i investigadors en aquesta àrea a escala espanyola. S'hi pot trobar informació sobre aquesta associació, el congrés anual que organitza i la revista que publica.

- **Associació Catalana d'Intel·ligència Artificial (ACIA):** http://www.acia.org/. És una organització d'àmbit català que té com a objectiu donar suport a la comunicació entre persones i organitzacions que treballen en aquest camp a Catalunya. Organitza un congrés anualment.

- **European Language resources Distribution Agency (ELDA):** http://www.elda.fr/catalog.html. És l'agencia d'ELRA (European Language Resources Association), que es va crear per identificar, classificar, col·leccionar, validar i distribuir recursos lingüístics. A l'apartat *Catalog* es troben enllaços a una àmplia gamma de recursos com lèxics, corpus i recursos relacionats amb les tecnologies de la parla.

- **Linguistic Data Consortium:** http://www.ldc.upenn.edu/. El Linguistic Data Consortium s'ocupa del desenvolupament de la tecnologia, la investigació i l'educació relacionades amb el llenguatge a través de la creació de recursos lingüístics de lliure distribució.

- **Elsnet:** http://www.elsnet.org/resources.html. Elsnet és una xarxa d'excel·lència europea dedicada a les tecnologies del llenguatge humà. En aquesta pàgina concreta s'incorpora una llista de recursos disponibles a través d'Elsnet.

- **Linguistics Computing Resources on the Internet:** http://www.sil.org/linguistics/computing.html. En aquest lloc web, que pertany al portal del Summer Institute of Linguistics (SIL), es presenta una llista general de recursos organitzats segons la temàtica.

2) Lèxics

- **WordNet:** http://wordnet.princeton.edu/. Base de dades lèxica de l'anglès que organitza el lèxic com una ontologia. Aquesta base de dades ha estat desenvolupada des de l'àmbit de la psicologia. Les paraules s'organitzen com a conjunt de formes que comparteixen un sentit, *synset*, i que es relaciona amb els altres *synsets* segons sigui un hiperònim o un hipònim.

- **EuroWordNet:** http://www.illc.uva.nl/EuroWordNet/. És la versió europea del Wordnet desenvolupada amb finançament de la UE per a llengües europees com l'italià, l'holandès, l'alemany, l'espanyol i el català, entre d'altres. No presenta una cobertura total i en algunes llengües no és de lliure disposició.

- **SenSem:** http://grial.uab.es/recursos.php. Lexicó de la llengua espanyola basat en les dades recopilades en el corpus del mateix nom. En l'actualitat s'està construint un recurs similar per el català.

- **FrameNet:** http://gemini.uab.es:9080/SFNsite. Lexicó de la llengua espanyola basat en la proposta de Framenet per a la llengua anglesa i creat a la Universitat de Berkeley.

- **ARIES Natural Language Tools:** http://www.mat.upm.es:80/~aries/description.html. Des d'aquest lloc es pot accedir a un conjunt d'eines que configuren una plataforma per a la representació i l'anàlisi del lèxic espanyol. Aquestes eines es poden integrar en aplicacions de processament de llenguatge natural.

3) Gramàtiques

- **PC-PATR:** http://www.sil.org/pcpatr/. D'aquest web es poden descarregar les distintes versions de l'analitzador sintàctic PC-PATR, basat en gramàtiques d'estructures de trets i en la unificació.

- **LFG Grammar Writer's Workbenck:** http://www2.parc.com/istl/groups/nltt/medley/. Aquest programa és una eina d'anàlisi basada en el formalisme de la gramàtica lexicofuncional de Kaplan i Bresnan (1982).

- **Minipar:** http://www.cs.ualberta.ca/~lindek/minipar.htm. És un analitzador per a l'anglès basat en la gramàtica de la teoria minimista.

4) Programes de processament

- **Signum:** http://www.lenguaje.com/. Aquest és el web d'una empresa que ha desenvolupat diversos programes per al processament de l'espanyol, com un conjugador verbal i un programa basat en la similitud fonètica, i algunes aplicacions, com un corrector ortogràfic, entre d'altres.

- **Natural Language Software Registry:** http://registry.dfki.de/. És un sumari dels programes de processament del llenguatge natural (i, excepcionalment, alguns recursos) estructurats i definits mitjançant descriptors.

- **Summer Institute of Linguistics:** http://www.sil.org/linguistics/computing.html. Dins de l'apartat Computers es proporciona una llista d'eines que es poden trobar a Internet i que són útils per organitzar i analitzar dades lingüístiques.

- **Interactive On Line CL Demos:** http://www.cl.uzh.ch/. Aquí trobareu una col·lecció d'enllaços a programes de demostració interactius sobretot per a la llengua anglesa i l'alemanya, com, per exemple, analitzadors morfològics i sintàctics, programes per al processament de corpus, eines per al processament de la parla, programes de resum i desambiguadors.

- **Institut Universitari de Lingüística Aplicada:** http://www.iula.upf.edu/. És un lloc web en què s'ofereix un recull de recursos desenvolupats a l'IULA, sobretot recursos lexicogràfics de terminologia i corpus multilingües.

- **TALP - Natural Language Research Group:** http://www.lsi.upc.es/~nlp/. En aquest web es poden consultar analitzadors morfològics i sintàctics.

5) Recursos morfològics

- **PC-Kimmo:** http://www.sil.org/pckimmo/. PC-KIMMO és un programa creat per Kimmo Koskenniemi que es pot instal·lar a ordinadors personals per analitzar i/o generar paraules utilitzant un model d'estructura de dos nivell, el lèxic profund i el superficial.

- **Grupo de Estructuras de Datos:** http://www.gedlc.ulpgc.es/investigacion/scogeme02/flexsus.htm. Aquest grup de la Universidad de Gran Canaria ha posat a disposició del públic una sèrie de recursos morfològics, com un generador verbal, un flexionador de noms i adjectius que permet generar augmentatius, diminutius, pejoratius, superlatius, etc. i un lematitzador.

- **Onoma:** http://www.onoma.es/. Accedint a aquesta pàgina es pot provar un conjugador de l'espanyol que es caracteritza perquè permet conjugar verbs inventats.

- **Verbix:** http://www.verbix.com/index.html. Aquesta eina és un conjugador de diverses llengües.

6) Tecnologies de la parla

- **Speech Processing Group - Universitat Politècnica de Catalunya:** http://gps-tsc.upc.es/. Aquest lloc web inclou informació sobre el grup d'investigació de tecnologia de la parla de la UPC. Es recomana la consulta dels apartats de demostracions i programari.

- **Emotional & Expressive Synthesized Speech:** http://xenia.media.mit.edu/~cahn/emot-speech.html. Aquest web ha estat elaborat per Janet Cahn del MIT Media Laboratory. Aquí es pot escoltar parla sintetitzada per a l'anglès. El més característic és que s'han sintetitzat exemples d'una mateixa frase interpretant diferents emocions.

- **The ASEL ModelTalker TTS System:** http://www.asel.udel.edu/speech/ModelTalker.html. Aquest és un altre exemple de conversor text-parla per a l'anglès. L'usuari pot escoltar directament del web alguns exemples de frases sintetitzades.

- **Examples of Synthesized Speech:** http://www.ims.uni-stuttgart.de/~moehler/synthspeech/examples.html. Des d'aquest web es pot accedir a diferents exemples de parla sintetitzada en diverses llengües.

- **Adaptive Technologies Research Center:** http://www.utoronto.ca/atrc/reference/tech/voicerecog.html. Aquest lloc web de la University of Toronto conté enllaços a adreces de programes comercials de reconeixement de veu. Inclou també informació sobre les especificacions tècniques d'aquests programes.

Bibliografia

Allen, J. (1987). *Natural Language Understanding*. Redwood City: Benjamin/Cummings.

Atserias, J.; Casas, B.; Comelles, E.; González, M.; Padró, Ll.; Padró, M. (2006). "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library". A: *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*. ELRA. Gènova.

Beale, S.; Nirenburg, S.; Mahesh, K. (1995). "Semantic Analysis in the Mikrokosmos Machine Translation Project". A: *Proceedings of the Second Symposium on Natural Language Processing* (SNLP-95) (pàg. 297-307). Bangkok: Kaser Sart University.

Castillo Holgado, A. (2001). "Riqueza y Patrimonio: el español en la sociedad del conocimiento". *Nueva revista de política, cultura y arte* (núm. 74, pàg. 148-150).

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

Charniak, E. (2000). "A maximum-entropy-inspired parser". A: *Proceedings of NAACL-2000* (pàg. 132-139).

Church, K.; Mercer, R. (1993). "Introduction to the Special Issue on Computational Linguistics Using Large Corpora". *Computational Linguistics* (vol. 19, núm. 1, pàg. 1-24).

Fellbaum, C. (ed.) (1998). *WordNet. An Electronic Lexical Database*. Cambridge: MIT.

Fernández, M. (1996). *Avances en lingüística aplicada*. Universidad de Santiago de Compostela.

Gazdar, G; Mellish, C. (1989). *Natural Language Processing in PROLOG*. Reading: Addison-Wesley.

Grishman, R. (1986). *Computational Linguistics*. Cambridge: Cambridge University Press. Traducció espanyola: *Introducción a la lingüística computacional* (1991). Madrid: Visor.

Grosz, B. (1996). *Survey of the State of the Art in Human Language Technology* [publicació en línia: http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html].

Hirschberg, J.; Grosz, B. (1992). "Intonational features of local and global discourse structure". A: *Proceedings of the Fifth DARPA Speech and Natural Language Workshop (pàg*. 441-446). Nova York: Arden House.

Jurafsky, D.; Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing:* A *Computational Linguistics and Speech Recognition*. Nova York: Prentice Hall.

Klavans, J.; Resnik, P. (editors) (1996). *The balancing act: combining symbolic and statistical approaches to languages*. Cambridge: MIT.

Llisterri, J.; Garrido Almiñana, J. M. (1998). "La ingenieria lingüística en España". A: *El español en el mundo. Anuario del Instituto Cervantes* [en línia: http://cvc.cervantes.es/lengua/anuario/anuario_98/]. Centro Virtual Cervantes.

Marcus, M. P.; Santorini, B.; Marcinkiewicz, M. A. (1994). "Building a large annotated corpus of English: The Penn Treebank". *Computational Linguistics* (núm. 19(2), pàg. 313-330).

Meya Llopart, M.; Huber, W. (1986). *Lingüística Computacional*. Barcelona: Ed. Teide.

Miller, G. A. (2009). "WordNet - About Us". *WordNet* [en línia: http://wordnet.princeton.edu/]. Princeton University.

Moreno Sandoval, A. (2008). "Panorama actual de la ingeniería lingüística". A: A. Alcina, E. Valero i E. Rambla (editors). *Terminología y Sociedad del* conocimiento (pàg. 100-115). Berlín: Peter Lang.

Payrató, Ll. (1997). *De professió, lingüista. Panorama de la lingüística aplicada*. Barcelona: Empúries.

Schank, R. (1975). *Conceptual Information Processing*. Nova York: Elsevier.

Smith, G. W. (1991). *Computers and Human Language*. Oxford: Oxford University Press.

Webber, B. L. (2001). "Computational perspectives on discourse and dialogue". A: A. D. Schirin, D. Tannen i H. Hamilton (editors). *The Handbook of Discourse Analysis* (pàg. 798-816). Malden, Oxford: Blackwell Publishers, Ltd.

Winograd, T. (1983). *Language as a Cognitive Process: Syntax*. Reading: Addison-Wesley.