Tecnologies de la llengua i les seves aplicacions

M. Antònia Martí Antonín
Montserrat Civit Torruella
Mariona Taulé Delor
CLiC-Centre de Llenguatge i Computació
{mtaule,amarti}@ub.edu civit@clic.fil.ub.es

Resumo:

A investigación en Lingüística Computacional e Procesamento da Lenguaje Natural deu lugar estes últimos anos ás denominadas Tecnoloxías da Linguaxe, cuxo obxectivo principal é o desenvolvemento de sistemas informáticos capaces de recoñeceren, comprenderen e xeraren linguaxe humana en todas as súas formas. Con esta finalidade, desenvolveuse unha serie de aplicacións, como a Tradución Automática, a Extracción e Recuperación da Información, a Clasificación de Documentos etc., que procesan a información para facilitaren o acceso, organización e transmisión do coñecemento que xera a chamada Sociedade da Información en que vivimos.

Como noutras disciplinas científicas, na área da Lingüística Computacional e do Procesamento da Linguaxe Natural pasouse dunha etapa inicial centrada na investigación básica de carácter experimental a outra en que se interaxe máis coa sociedade e, por tanto, máis interesada na creación de produtos e aplicacións que resolvan problemas reais. Isto significa desenvolver sistemas e recursos capaces de analizaren a linguaxe sen restricións, isto é, que ofrezan unha ampla cobertura lingüística.

Neste artigo preséntase de xeito introdutorio os recursos (lingüísticos) e as aplicacións máis características que se desenvolven actualmente no marco das Tecnoloxías da Linguaxe. En concreto, salientaremos dos recursos necesarios os analizadores e desambiguadores morfolóxicos e sintácticos, os lexicóns computacionais e os corpus lingüísticos, nomeadamente os etiquetados. Canto ás aplicacións, centrarémonos básicamente na Recuperación e Extracción da Información e na Tradución Automática.

Palabras chave:

Procesamento da Linguaxe Natural, Lingüística Computacional, Recursos Lingüísticos.

Abstract:

In the last years, research on Computational Linguistics and Natural Language Processing has led to Language Technologies, whose main goal is to develop computer systems capable to recognize, understand and generate human languages in all their forms. For this purpose, several applications have been developed, such as Machine

Translation, Information Retrieval and Information Extraction or Document Classification. These applications process the language in order to ease access to knowledge, its organization or its transmission, activities needed by our Information Society.

As in other disciplines, Computational Linguistics and Natural Language Processing have gone from a first period of basic, experimental research to another in which new products and real applications have to be created, in order to solve interaction problems. This means that we need to develop systems and resources capable to deal with unrestricted language, that is, broad-coverage systems and resources. This paper presents an introduction to linguistics resources as well as the main applications being developed nowadays in the Language Technologies framework. More concretely, it emphasizes morphological analyzers, taggers, syntactic parsers, computational lexicons and linguistic annotated corpora. As for applications, stress is laid on Information Retrieval, Information Extraction and Machine Translation.

Key words:

Natural Language Processing. Computational Linguistics. Linguistic Resources.

1. Introducció

El llenguatge ha estat un instrument decisiu en el procés d'humanització que ens ha permès associar-nos, intercanviar informació i transmetre coneixement. Llenguatge i Societat són, per tant, dos conceptes que estan estretament units i no s'entenen ni s'expliquen l'un sense l'altre. En l'evolució de la cultura, l'aparició de l'escriptura constitueix una fita cabdal ja que permet emmagatzemar informació i disposar d'una memòria perdurable enfront del caràcter efímer de la llengua oral. En aquesta línia, la invenció de la impremta constitueix una innovació podríem dir que revolucionària perquè permet la difusió a gran escala del coneixement humà, que influeix decisivament en el procés d'alfabetització i, conseqüentment, a la democratització de la cultura.

A finals del segle XX, la codificació digital de la informació constitueix una nova revolució pel que fa a l'emmagatzemament, organització i transmissió de dades textuals. L'ús generalitzat dels ordinadors personals i, sobretot, l'aparició d'Internet, ha propiciat un creixement exponencial de la quantitat d'informació disponible. Aquestes facilitats en les comunicacions han contribuït a la creació d'un nou nivell de realitat, la realitat virtual, que ens permet realitzar una àmplia gamma d'interaccions. En aquesta nova realitat les distàncies desapareixen de manera que la nostra relació amb l'entorn esdevé global.

La situació actual és, però, d'alguna manera paradoxal: d'una banda, disposem de gran quantitat d'informació susceptible de ser consultada, adquirida i difosa i,

d'una altra, la quantitat d'informació és tan gran i de naturalesa tan variable que es fa molt difícil accedir-hi de manera adient. El resultat ha estat que ha canviat substancialment el concepte mateix del que es considera informació i com tractar-la: no està millor informat qui té més dades, sinó qui disposa dels mitjans necessaris per obtenir específicament aquelles dades que necessita o cerca, aquelles dades que són més apropiades per als seus interessos. És més, el que ara es vol obtenir és coneixement, és a dir, dades organitzades, estructurades i relacionades.

El concepte de Societat de la Informació, és a dir, una societat que es recolza en la producció, organització i transmissió digital de la informació, ha d'incorporar al bé comú que és el llenguatge el desenvolupament de tècniques que facilitin la transmissió del saber, l'organització, interrelació i contextualització de les dades i les parcel·les de coneixement disperses.

L'entorn comunicatiu que deriva de l'aplicació de les noves tecnologies de la llengua planteja reptes fins ara inexistents. Cal dissenyar sistemes informàtics que permetin l'accés a la informació de manera eficaç, ràpida i senzilla. És amb aquesta finalitat que, entre d'altres, les aplicacions de tecnologia lingüística tenen com a objectiu:

- la classificació de documents de manera que es faciliti la seva posterior recuperació;
- el resum automàtic per tal de simplificar-ne la consulta;
- la traducció multilingüe per poder accedir a la informació de documents en diferents llengües;
- el disseny d'interfícies que permetin l'accés a les dades organitzades i classificades;
- l'accés a la informació de manera precisa, com és el cas dels sistemes pregunta-resposta;
- etc.

És desitjable també que s'integrin en un mateix entorn la veu, la imatge i la llengua escrita i que es tingui en compte la varietat de llengües en què es pot trobar la informació, tant per obtenir-la com per poder recuperar els documents que la contenen.

Resoldre aquests problemes exigeix nous plantejaments en el tractament de determinats temes amb llarga tradició en el marc de les tecnologies de la llengua com són la traducció automàtica i la classificació documental, i alhora obre portes al desenvolupament de noves aplicacions com les que acabem d'esmentar. La revolució tecnològica que ha significat l'aparició dels nous sistemes de

comunicació està provocant a més un canvi en profunditat en les professions i sectors productius relacionats amb el llenguatge, com són la traducció, l'edició -en especial de diccionaris i enciclopèdies- i l'ensenyament, que exigeixen una readaptació en profunditat dels professionals d'aquestes àrees.

La possibilitat de dur a terme aquestes noves aplicacions o d'innovar en àrees de coneixement ja existents requereix el desenvolupament de recursos d'enginyeria lingüística que són a la base de qualsevol tractament informàtic del llenguatge humà. Es tracta de sistemes d'anàlisi morfològica, de mòduls de desambiguació morfosintàctica, de bases de coneixement lèxic, de lexicons computacionals, de gramàtiques, d'analitzadors, etc., que són específiques de cada llengua. El desenvolupament d'aquest tipus de recursos requereix de la confluència de diferents agents, públics i privats, atesa la seva complexitat i l'elevat cost de desenvolupament constitueixen el nucli de les anomenades Indústries de la Llengua.

1.1. Les Indústries de la llengua

El Processament del Llenguatge i les Tecnologies de la parla constitueixen la base de l'enginyeria lingüística: l'aplicació dels coneixements sobre la llengua al desenvolupament de sistemes informàtics que puguin reconèixer, comprendre, interpretar i generar llenguatge humà en totes les seves formes.

Encara que el tractament computacional del llenguatge es va iniciar a mitjan dels anys quaranta, no ha estat fins a la darrera dècada del segle XX, als anys noranta, que s'ha iniciat el desenvolupament de sistemes de tractament del llenguatge a gran escala. Aquest canvi d'enfocament ha estat determinat fonamentalment per dos factors bàsics: la difusió de l'ordinador personal, el PC, com a eina habitual per a codificar la informació i l'aparició d'Internet com a mitjà per a transmetre i publicitar la informació. Aquests dos factors han determinat un creixement exponencial de la informació en suport electrònic i, conseqüentment, la necessitat de desenvolupar sistemes informàtics per a accedir-hi de manera intel·ligent.

De la mateixa manera que s'ha observat en altres ciències i disciplines, en l'àrea de la Lingüística Computacional i del Processament del Llenguatge s'ha passat d'una primera etapa centrada en la recerca bàsica i en el desenvolupament de sistemes de laboratori de caràcter experimental, a una altra en què l'objectiu és constituir-se en una ciència que interactua amb la societat i que proporciona una base per a la creació de productes de caràcter aplicat.

Recentment s'ha encunyat el terme d'"Indústries de la Llengua" per a designar el complex entramat de productes i serveis que es desenvolupen al voltant del llenguatge: l'edició en suport electrònic, els correctors automàtics, el tractament de

la informació a Internet, el desenvolupament de bancs de dades textuals, etc. precisen de recursos i aplicacions d'Enginyeria Lingüística que facilitin la gestió i l'accés de la informació. En aquests darrers anys, ha aparegut, per tant, una notable demanda d'aquest tipus de serveis i productes lingüístics impensable fa només vint anys, i que ha significat l'aparició de noves professions, necessitats i productes que tenen com a base el tractament informàtic del llenguatge.

Per contra, aquesta nova situació pot agreujar els problemes que tenen plantejats les llengües minoritàries en la mesura que no disposen dels recursos lingüístics necessaris per a participar de ple en la Societat de la Informació. Els criteris de rendibilitat sens dubte decanten l'interès en el desenvolupament de recursos d'enginyeria lingüística per a llengües amb una àmplia difusió i prestigi social i científic. Així, la Societat de la Informació pot actuar com a factor negatiu en el sentit de desequilibrar més la balança a favor d'unes determinades llengües amb una àmplia difusió i amb una presència predominant a Internet, com l'anglès, i a deixar cada cop més en un segon pla les llengües en una situació d'ús més restringida.

Aquest desequilibri es manifesta de maneres molt diverses. D'una banda, s'està agreujant les diferències entre països amb una alta implantació de les noves tecnologies i països en què pràcticament no s'han alterat les maneres de comunicar-se i transmetre la informació. D'una altra, en els països amb una presència rellevant de les noves tecnologies s'està decantant la balança a favor de l'ús de les llengües predominants, per damunt de llengües minoritzades. En les interaccions comunicatives digitals primen els criteris de rendibilitat, cost i impacte de la informació: es tria la llengua que permetrà arribar a més gent ja que els límits d'espai han desaparegut i el missatge pot arribar arreu.

A partir d'ara, entre els factors que es prenen en consideració a l'hora d'avaluar l'estatus d'ús d'una llengua, caldrà també tenir en compte la seva presència en el marc de les tecnologies lingüístiques i l'existència tant de recursos computacionals com d'aplicacions per al seu tractament automàtic.

Si no es vol agreujar la situació en què es troben les llengües minoritzades, s'hauran de prendre mesures que rectifiquin la tendència a la substitució lingüística en determinats contextos d'ús, la qual cosa implica la complicitat de tots els agents implicats que han de facilitar els mitjans econòmics i estratègics per a redreçar aquesta situació.

1.2. Relacions entre la Lingüística i la Lingüística Computacional

La Lingüística Computacional va establir des de la seva aparició una interacció dinàmica amb la Lingüística teòrica que les ha enriquides mútuament. D'una

banda, la Lingüística Computacional va adoptar des d'un bon començament tot l'aparell formal que aportava la Gramàtica Generativa, que incloïa, entre d'altres, el concepte de regles de reescriptura o de producció i el de transformacions; la Lingüística teòrica, d'altra banda, ha obtingut, gràcies a la Lingüística Computacional, la possibilitat d'implementar els models teòrics i desenvolupar un aparell experimental que permet posar-los a prova.

En un terreny més general, el fet de disposar de grans quantitats de corpus en suport electrònic ha permès que els lingüistes tinguin accés a mostres d'ús real de la llengua. Gràcies a les tecnologies de la llengua es poden anotar corpus a gran escala i de manera automàtica a diferents nivells- morfològic, sintàctic i semàntic. A partir dels corpus anotats s'obté informació precisa sobre el comportament de les llengües que permet verificar hipòtesis des de la lingüística i millorar els lexicons i les gramàtiques computacionals. A continuació presentem amb un cert detall el que ha estat la interacció entre la Lingüística i la Lingüística Computacional així com l'impacte de la Lingüística de Corpus en ambdues disciplines.

La publicació el 1957 del llibre *Syntactic Structures* de N. Chomsky i, posteriorment, el 1965, d'*Aspects of one Theory of Syntax* van constituir una aportació de gran transcendència en el desenvolupament de la Lingüística, que va repercutir significativament en altres disciplines com la Filosofia, la Psicologia i la Lingüística Computacional. Chomsky demostrava la insuficiència dels autòmats d'estats finits i de les gramàtiques independents del context com a models generadors de frases d'una llengua natural, i proposava una teoria del llenguatge basada en dos nivells de representació, l'estructura profunda i la superficial, i en dos tipus d'objectes formals: les regles de reescriptura i les transformacions. Per primera vegada una teoria lingüística tenia com a objectiu desenvolupar un aparell formal que permetia descriure i generar les frases de les llengües naturals. Aquesta nova concepció de la gramàtica encaixava perfectament amb els requisits de formalització exigits pels sistemes de processament del llenguatge, que són inherentment formals, ja que els algorismes operen sobre dades discretes de forma que aquells aspectes de les teories que no són formalment expressables no es poden processar.

A partir d'ara i en les diferents formulacions que aniran apareixent, l'objectiu de la teoria lingüística serà caracteritzar totes les possibles llengües naturals i, a la vegada, proporcionar l'aparell formal que permeti la descripció de les llengües en termes de la teoria. Explicar un fet és preveure de quina manera i sota quines condicions es produirà. Així, una gramàtica generativa ha de tenir codificat tot el coneixement lingüístic que permetrà produir totes i només aquelles frases que són gramaticals en una llengua.

Aquests avenços en el terreny lingüístic van tenir una repercussió immediata en el desenvolupament de la Lingüística Computacional a partir de la dècada dels setanta i sobretot als anys vuitanta, en què van començar a aparèixer formalismes i teories sobre el llenguatge amb una clara voluntat de ser processables. Des d'aleshores, l'ús de regles d'estructura de frase, de transformacions, de condicions de bona formació, etc. han estat incorporades en els models computacionals de les llengües naturals. S'havia demostrat en les dues dècades anteriors, en els sistemes de laboratori, que era factible solucionar determinats problemes de processament del llenguatge i ara calia fer un pas endavant, desenvolupar aplicacions reals. Per tal d'assolir aquest objectiu era necessari promoure el desenvolupament de recursos bàsics a gran escala, en especial gramàtiques, lexicons i bases de dades textuals, per al màxim nombre de llengües possible. Com a resposta a aquestes necessitats, al Regne Unit es promou el programa Alvey (Oakley i Owen 1990) i a la Unió Europea s'afavoreix la investigació bàsica i el desenvolupament de recursos lingüístics a gran escala.

Els anys vuitanta es caracteritzen bàsicament per la irrupció de les teories lingüístiques en l'àmbit dels sistemes de processament del llenguatge i per la voluntat de superar les limitacions imposades a partir del concepte de subllenguatge, que limitaven les aplicacions computacionals al tractament de dominis conceptuals i lingüístics restringits (Kitterdge i Lehrberger 1982). El que es pretén a partir d'aquest moment és la creació de sistemes que es caracteritzen per ser capaços de processar qualsevol tipus de text sense cap mena de restricció: les gramàtiques computacionals han de ser de gran abast i tractar una gran varietat de fenòmens lingüístics.

L'exigència de rigor en la formalització del llenguatge és cada cop més gran i els formalismes i teories gramaticals que apareixen al llarg dels anys vuitanta tenen una clara intenció de ser processables computacionalment. Es pot dir que a partir d'aquests moments el tractament computacional és un nou requisit que garanteix que una teoria disposa d'un aparell formal consistent i suficientment expressiu. En aquesta línia, es desenvolupen models lingüístics que tenen com a objectiu captar al màxim les regularitats del llenguatge i així fer generalitzacions lingüístiques el més apropiades possible.

Com a resultat d'aquest enfocament es desenvolupen els formalismes d'unificació o gramàtiques d'unificació, models teòrics o simples anotacions formals, de caràcter marcadament declaratiu, que tenen en comú l'ús d'estructures de trets per a la descripció lingüística i que apliquen l'operació d'unificació per a l'estructuració dels objectes. Entre els primers destaquen la Gramàtica Funcional Lèxica (*Lexical Functional Grammar*, LFG, Kaplan i Bresnan 1982), la Gramàtica d'Estructura Sintagmàtica Generalitzada (*Generalized Phrase Structure Grammar*, GPSG, Gazdar et al. 1985) i la Gramàtica d'Estructura Sintagmàtica Regida pel

Nucli (*Head-Driven Phrase Structure Grammar*, HPSG, Pollard i Sag 1987); entre els segons, cal esmentar el formalisme PATR-II (Shieber 1986), la Gramàtica Funcional d'Unificació (*Functional Unification Grammar*, FUG, Kay 1985), i la Gramàtica d'Adjunció d'Arbres (*Tree Adjoining Grammar*, TAG, Joshi 1984)[1].

La construcció de gramàtiques generals de la llengua planteja el problema de com obtenir informació objectiva i a gran escala, ja que les intuïcions del lingüista no són suficients per a un treball d'aquesta magnitud. En aquest sentit, l'adquisició de coneixement lingüístic es constitueix com una línia de recerca clau per al futur desenvolupament de la Lingüística Computacional per tal que aquesta pugui assolir els objectius que té traçats: la superació de les limitacions imposades pels subllenguatges, la construcció de gramàtiques d'àmplia cobertura i la incorporació de coneixement lingüístic per a la resolució de l'ambigüitat. És en aquest punt que els corpus textuals en suport electrònic adquireixen una importància cabdal ja que esdevenen la principal font d'informació per a inferir coneixement lingüístic, bàsicament per veure de quina manera s'usa una llengua, quin tipus de construccions sintàctiques predominen, quina mena de vocabulari utilitza un determinat domini, en definitiva, per fornir la informació necessària per a la construcció de gramàtiques i lexicons computacionals.

1.2.1. La Lingüística de corpus

Un corpus és una mostra d'una llengua que habitualment s'ha construït a partir d'una selecció de textos realitzada segons uns determinats criteris i amb un objectiu concret. Entre els criteris de constitució d'un corpus podem citar: la recopilació de l'obra d'un autor, la representació de la llengua en un moment determinat de la seva evolució, l'exemplificació de determinats fets de la llengua, la selecció de material sobre un domini temàtic específic, etc. Pel que fa als objectius, com acabem d'esmentar en l'apartat anterior, es tracta fonamentalment d'obtenir dades d'ús de les llengües ja sigui tant per a la seva anàlisi lingüística com computacional.

L'existència d'una disciplina específica, la Lingüística de Corpus, que tracta sobre la constitució i processament d'aquest tipus d'objectes, es justifica pel fet de requerir d'una metodologia que difereix en molts sentits de l'emprada per al processament de petites mostres de la llengua. La tipologia de corpus, els criteris per construir-los, les tècniques de processament, la seva finalitat, etc. constitueixen línies de treball clarament diferenciades.

En el seu origen, la Lingüística de Corpus consisteix en l'estudi de les llengües a partir d'exemples d'ús. Els estudis lingüístics basats en corpus no són nous ni tenen

[1] Per a una informació de síntesi sobre formalismes d'unificació es pot consultar A. Moreno Sandoval (2001), M. Taulé Delor i M.A. Martí Antonín (2002) i S. Shieber (1986).

a veure directament amb la Lingüística Computacional: els primers estructuralistes basaven els seus estudis en l'anàlisi de textos. En paraules de Harris (1993): "The approach began … with a large collection of recorded utterances from some language, a corpus. The corpus was subjected to a clear, stepwise, bottom-up strategy of analysis".

Ara bé, des de fa uns anys, *Corpus* és sinònim de *Corpus en suport electrònic* i *Lingüística de Corpus* ho és de *Lingüística de Corpus en suport electrònic*. L'aparició i l'ús de l'ordinador permet, d'una banda, buscar, recuperar, ordenar i fer càlculs sobre les dades de manera ràpida i eficaç; d'una altra, permet el maneig d'un volum de dades inimaginable fa poc més d'una dècada. La Lingüística de Corpus tal i com s'entén en l'actualitat, està relacionada no tan sols amb l'anàlisi i interpretació de la llengua sinó també amb les tècniques computacionals i la metodologia per a l'anàlisi de textos. La *Lingüística de Corpus en suport electrònic*, segons Leech (1992), focalitza la seva atenció en l'actuació lingüística més que no pas en la competència, en la descripció lingüística, en els models de llenguatge tan qualitatius com quantitatius i en un enfocament de la recerca empirista enfront de l'enfocament racionalista.

Quin és l'interès des d'un punt de vista tant teòric com aplicat, de la Lingüística de Corpus? En primer lloc, cal assenyalar que l'ús de dades externes al parlant, resultat de la producció lingüística, presenta avantatges evidents sobre la introspecció: es tracta de dades directament observables, mentre que la introspecció és subjectiva i opinable. En segon lloc, els corpus són una font d'informació quantitativa important. La freqüència d'aparició dels elements lingüístics no pot recuperar-se per introspecció i si una paraula o una determinada estructura no apareix en un corpus representatiu d'una llengua pot ser una informació significativa, ja que pot qüestionar judicis acceptats per inèrcia. Finalment, els corpus fan possible la verificació objectiva de resultats, especialment quan es tracta de comprovar el bon funcionament d'eines per al processament del llenguatge, però també quan es tracta d'aportar proves per a corroborar les hipòtesis d'una teoria[2].

Així, el desenvolupament de metodologies basades en l'anàlisi de corpus lingüístics centren l'interès dels estudis entorn del Processament del llenguatge perquè (a) constitueixen una font d'informació valuosa per a la creació de diccionaris i lexicons tant per a l'ús humà com computacional; (b) perquè són una font d'informació fonamental per al desenvolupament de gramàtiques computacionals; (c) perquè degudament anotats a nivell morfosintàctic o semàntic, s'usen en els processos d'aprenentatge automàtic per a construir desambiguadors

[2] Vegeu per a més informació sobre el tema: Leech (1997), McEnery i Wilson (1996), Ooi (1998) i Civit (2003).

morfològics, sintàctics o semàntics; (d) perquè constitueixen un banc de proves valuós per a evaluar la qualitat dels sistemes d'anàlisi del llenguatge; i, finalment, (e) perquè el propi estudi del llenguatge es pot enriquir gràcies a l'aportació de l'anàlisi de dades textuals.

Sens dubte, una de les aplicacions més exteses i conegudes dels corpus és el seu ús com a font d'informació per a delimitar els sentits dels mots i obtenir exemples, extreure locucions i col·locacions, en definitiva, per a l'elaboració de diccionaris d'ús comú. L'exemple més conegut i pioner va ser el diccionari Collins-Cobuild (Sinclair, 1987), que es va construir íntegrament a partir de la consulta dels exemples existents en un corpus de 6 milions de paraules creat amb aquest fi per la pròpia editorial i la Universitat de Birmingham. Actualment no es concep l'elaboració d'un diccionari que informi sobre l'ús de la llengua sinó és sobre la base d'un corpus representatiu. Els corpus que ara s'utilitzen, però, són de dimensió molt superior, entre els 100 i 200 milions de paraules i normalment s'han etiquetat tant a nivell morfosintàctic com semàntic.

2. Recursos, eines i processos bàsics de Tecnologia Lingüística

Els darrers anys, les línies que han marcat l'evolució de la Lingüística Computacional han estat el desenvolupament, l'estandardització i la compatibilització de recursos lingüístics amb l'objectiu de dotar a les llengües de la infraestructura necessària per a participar en la Societat de la Informació.

Les aplicacions computacionals que tracten informació textual han de fer front a una sèrie de problemes que són característics i inherents a les llengües: la riquesa morfològica i semàntica i l'ambigüitat. Els analitzadors morfològics i sintàctics han de ser capaços de construir representacions formals del contingut dels textos en termes de la seva estructura: han d'assignar la categoria correcta a cada paraula, identificar els constituents i les funcions de les oracions i construir-ne una representació. Les bases de coneixement lexicosemàntic han de permetre associar el sentit correcte a cadascuna de les paraules d'un text i així fer possible la interpretació semàntica de les oracions. Tant en un cas com en l'altre són necessaris programes de desambiguació que seleccionin la categoria morfosintàctica correcta, l'estructura sintàctica adequada i el sentit apropiat de cada ocurrència en el seu context.

Des d'un punt de vista informàtic, les tècniques que es fan servir, i que són a la base dels programes de processament del llenguatge, es basen fonamentalment en dos tipus de coneixement: o bé es recolzen en coneixement lingüístic, normalment expressat en forma d'estructures de dades declaratives i regles, o bé es basen en l'estadística a partir de mostres molt extenses de la llengua, és a dir,

de corpus en brut o bé etiquetats. Els sistemes basats en l'estadística construeixen el seu model de llenguatge inferint-lo a partir dels corpus; en els sistemes basats en el coneixement, en canvi, el model de llenguatge ha estat explicitat prèviament pels lingüistes en forma de gramàtiques, lexicons, bases de coneixement, etc.

En aquest apartat ens centrarem en la descripció dels recursos d'enginyeria lingüística més representatius i que s'usen en el sistemes de processament basats en el coneixement. Alguns, com ara els analitzadors morfològics són a la base de gairebé totes les aplicacions, d'altres, com les xarxes semàntiques tenen actualment un àmbit aplicatiu molt més restringit i s'usen fonamentalment en programes de recerca.

2.1. L'anàlisi morfològica

El component morfològic dels sistemes de processament del llenguatge actua en les primeres fases de l'anàlisi i du a terme una sèrie de tasques que són fonamentals per als processos posteriors d'anàlisi sintàctica i d'interpretació semàntica. És per això, que li dedicarem una especial atenció.

Els analitzadors morfològics reben com a input el text enregistrat en suport magnètic i el resultat que generen és el text segmentat en unitats lèxiques i categoritzat morfològicament. Per tal d'obtenir aquest resultat, l'analitzador morfològic ha de realitzar una sèrie de tasques com són la segmentació del text en unitats lèxiques; la seva lematització, és a dir, l'associació a cadascuna d'aquestes unitats del seu lema i, finalment, l'assignació a cada unitat d'una o més interpretacions morfològiques especificades mitjançant la categoria El procés d'anàlisi morfològica permet, doncs, relacionar totes les variants flexives d'un mot amb el seu lema i assignar-los informació referent a la categoria i als trets morfològics.

Per exemple, com a resultat d'un procés d'anàlisi morofològica de la paraula *cases*, podem obtenir la següent informació (Figura 1):

Forma	Lema	Informació morfosintàctica
cases	casa	NCFP (nom comú, femení, plural)
	casar	VM2IPS (verb no auxiliar, 2a. persona, indicatiu, present, singular)

Figura 1.

on la paraula *cases* rebrà dues interpretacions possibles, una com a nom i l'altra com a verb.

Pel que fa al disseny d'un analitzador, aquest està determinat per diferents factors, entre els quals cal assenyalar:

- les característiques morfològiques de la llengua;
- els tipus de processos morfològics que ha de tractar: flexió, derivació i composició;
- les unitats que es volen identificar: mots aïllats, lexies, temps compostos, perífrasis, etc.

Un analitzador morfològic adient per processar el gallec o el català no necessàriament ha de ser vàlid per processar altres llengües, com per exemple, l'alemany o el turc, tot i que quan es dissenya un analitzador es procura que sigui el més independent possible de qualsevol llengua concreta.

Els analitzadors morfològics es basen normalment en l'estratègia de declarar, d'una banda, les arrels dels mots i els sufixos i, de l'altra, de definir les regles de combinatòria de les unes amb els altres. Es tracta, en darrer terme, de no haver d'introduir una a una totes les variants flexives dels lemes, sinó d'obtenir-les per combinació dels seus morfemes constitutius. Per a llengües poc flexives, com és el cas de l'anglès, una solució habitual per al tractament morfològic és construir un lexicó de formes amb la informació morfològica associada (Vegeu Figura 2).

Lexicó de formes		Expansió del codi
love	V00P, NCS	V00P = verb (V), present (P), no té persona ni nombre (00)
loves	V3SP	NCS = nom (N), comú (C), singular (S)
loved	V00T	V3SP = verb (V), (3) persona, singular (S), present (P)
table	NCS	V00T = verb (V), passat (T), no té persona ni nombre (00)
tables	NCP	NCP = nom (N), comú (C), plural (P)
green	AQ	AQ = adjectiu (A), qualificatiu (Q)

Figura 2.

En el cas de les llengües romàniques, com el gallec o el català, amb una gran riquesa flexiva, en canvi, és recomanable disposar d'un sistema computacional que permeti disposar d'aquest tipus d'informació sense haver de declarar una a una totes les formes de la llengua i la informació associada corresponent. Són precisament els analitzadors morfològics els encarregats de dur a terme aquesta tasca. Podem distingir dos enfocaments possibles en l'estratègia d'anàlisi:

- sistemes d'anàlisi morfològica que actuen a la vegada que analitzen el text: la combinatòria d'arrels i sufixos es realitza en el mateix moment de l'anàlisi;
- sistemes que prèviament generen totes les formes possibles i analitzen el text mitjançant la comparació de les formes del text amb les formes generades automàticament: quan hi ha identitat entre la forma del text i alguna de les formes generades, s'associa a la forma del text la informació continguda en el formari.

Els analitzadors morfològics es basen en la definició de regles que expressen les combinacions possibles d'arrels i sufixos. La construcció d'un analitzador morfològic implica, per tant, definir les classes flexives d'una llengua: cada classe està constituïda per aquelles paraules que presenten el mateix comportament flexiu. Les paraules han de ser analitzades en funció de l'arrel i del sufix (o sufixos) flexiu(s) que admeten. De manera que cada classe d'arrels té un codi que la identifica i els sufixos, al seu torn, s'agrupen en classes segons la categoria que denoten i la classe d'arrels amb què combinen.

Cal tenir en compte, però, que els analitzadors morfològics analitzen textos escrits i els signes ortogràfics no tenen una correspondència exacta amb els sons de la llengua. Així, per al català tenim que els verbs *saltar*, *pelar*, *tallar*, etc. pertanyen a la classe dels verbs regulars de la primera conjugació, mentre que *caçar*, *pegar*, *pecar*, etc. tot i ser igualment verbs regulars de la primera conjugació, presenten canvis ortogràfics i, per tant, s'han de tractar de manera diferent, altrament podríem obtenir com acceptables formes com *caçes (la forma correcta és *caces*) o *menjem (la forma correcta és *mengem*).

Pel que fa a les tècniques d'anàlisi, els autòmats d'estats finits s'han mostrat especialment adients per a la generació i l'anàlisi automàtica de paraules ja que permeten implementar tant les alternances morfoortogràfiques (*cosir*/*cuso*) com les combinacions d'arrels i sufixos. En aquest marc, hi ha sistemes que implementen un model morfològic de dos nivells (Koskeniemmi 1983), el lèxic i el superficial, que permet tractar les irregularitats sistemàtiques del tipus *caçar*/*caces* a partir d'una variant ortogràfica que es considera bàsica, mentre que d'altres (Martí 1988) es basen en un model d'un únic nivell que obliga a declarar en el lexicó d'arrels les dues variants ortogràfiques: *caç-* i *cac-*.

Des d'una perspectiva aplicada, la problemàtica de l'anàlisi morfològica és un problema resolt. Existeixen analitzadors morfològics per a llengües molts diverses amb una qualitat del 100% en els resultats.

2.2. Desambiguació morfològica

Els sistemes d'anàlisi sintàctica reben com a *input* els textos segmentats i categoritzats morfològicament. El resultat de l'anàlisi morfològica sol ser ambigu, és a dir, que una mateixa paraula pot rebre més d'una interpretació ja que l'anàlisi morfològica s'aplica sense tenir en compte el context. L'anàlisi sintàctica es pot complicar de manera crítica si el text que ha de processar és morfològicament ambigu perquè s'incrementarà notablement el nombre d'anàlisis possibles. Així, per tal de procedir a l'anàlisi sintàctica o, simplement, per tal d'obtenir una anàlisi morfològica completa i més acurada, cal aplicar un procés de desambiguació en el qual cada paraula rebi només una única interpretació, aquella que li correspon en funció del context en què es troba.

Els desambiguadors morfosintàctics (en anglès, *taggers*) poden basar-se en el coneixement lingüístic o bé en dades estadístiques. En el primer cas, es té en compte el context immediat del mot, les dues o tres paraules que el precedeixen i el segueixen a més d'un conjunt de regles de desambiguació. Un exemple és el cas de la paraula *que*, que pot tenir dues anàlisis possibles: pronom o conjunció. El desambiguador morfosintàctic podria formular les regles següents per determinar en cada cas quina és la categoria més apropiada:

R1: si la paraula que la precedeix és un nom, aleshores és un pronom.

R2: si la paraula que la precedeix és un verb, aleshores és una conjunció

Els desambiguadors que es basen en criteris estadístics resolen l'ambigüitat en base a la solució més freqüent en un determinat context. Aquests sistemes han de disposar prèviament de corpus lingüístics correctament analitzats i desambiguats a partir dels quals poder 'aprendre' les solucions més freqüents donat un context. S'ha comprovat, per exemple, que la paraula *la* és gairebé sempre article, algunes vegades pronom i molt rarament el substantiu *la* (és a dir, la nota musical). Davant de l'ambigüitat els sistemes basats en el coneixement lingüístic decidiran la categoria en funció de les categories veïnes, mentre que els sistemes basats en l'estadística la decidiran en funció de l'opció més freqüent en el context en què es troba a partir de les ocurrències de la paraula *la* en textos prèviament desambiguats a mà. Hi ha també sistemes mixtos que combinen ambdós tipus de coneixement (Màrquez et al. 2001).

Com hem vist, el component morfològic associa informació sobre la categoria i sobre els trets morfològics a cada una de les formes d'un text i, mitjançant l'aplicació de tècniques de desambiguació, s'associa a cada forma un únic lema i una única descripció morfològica. Un cop analitzat, lematitzat i desambiguat, el text es pot sotmetre a un processament d'anàlisi sintàctica i d'interpretació

semàntica o bé es pot utilitzar com a font d'informació per a l'extracció d'informació morfosintàctica i lèxica. Així, es poden obtenir:

- freqüències de lemes i formes;
- freqüències de categories;
- freqüències de seqüències de categories;
- combinacions d'un lema amb determinades categories;
- frases en què apareix un determinat lema o forma;
- etc.

Ara bé, la informació que abasta un analitzador morfològic és limitada. Cal que els mots tinguin també associada informació sobre el seu comportament sintàctic i el seu tipus semàntic per tal de fer possibles altres tècniques de processament del text. Els recursos computacionals que forneixen aquest altre tipus d'informació entorn del mot són els lexicons computacionals i les xarxes semàntiques.

2.3. Lexicons computacionals

La lexicografia computacional té com a objectiu la creació de lexicons que continguin la informació necessària per als sistemes de processament del llenguatge natural i que a més siguin tractables computacionalment. Ha de donar solucions als problemes que planteja la complexitat intrínseca del lèxic, tant pel volum de dades que ha de tractar com per la seva varietat i complexitat.

Des del punt de vista de l'anàlisi computacional, el lèxic constitueix un dels principals problemes que han de resoldre els sistemes de processament del llenguatge natural, ja que de les aplicacions "pilot" dels anys setanta i vuitanta s'ha passat al disseny de sistemes "reals" que tracten les llengües naturals sense cap mena de restricció i això implica poder tractar qualsevol paraula en qualsevol de les seves formes i accepcions possibles.

En els sistemes de processament del llenguatge el component lèxic és una peça clau perquè, d'una banda, permet associar informació a les formes tractades en el component morfològic i, d'una altra, guia el procés d'anàlisi sintàctica i d'interpretació semàntica.

Les unitats de l'anàlisi lexicogràfica són els sentits, és a dir, cadascun dels diferents significats que pot tenir un mot. El fet més habitual és que els mots de les llengües naturals siguin polisèmics, és a dir, que tinguin més d'un significat. El problema, tant de la lexicografia tradicional com dels lexicons computacionals, és delimitar els diferents sentits d'un mot. Cada cop s'accepta més la idea que la identificació

dels sentits està estretament relacionada amb l'objectiu aplicatiu d'un lexicó: els sentits d'un mot variaran si estem construint un lexicó monolingüe o bilingüe, si l'objectiu és construir un sistema de traducció automàtica, d'extracció d'informació o de resum automàtic de textos. De la mateixa manera, en el terreny de la lexicografia tradicional s'observa que les accepcions d'un mot no són les mateixes en un diccionari monolingüe que en un bilingüe, en un diccionari orientat a l'ensenyament de segones llengües o a la consulta general (Martí 2003).

La unitat sobre la qual s'organitzen aquest tipus de recursos és el lema o bé els sentits d'un mot. De fet, la informació que conté un lema o una accepció en un lexicó computacional es considera vàlida per a totes les seves variants flexives. Els lemes són, per tant, les unitats que permeten relacionar les formes d'un text processat morfològicament amb altres dispositius computacionals.

Existeix un cert acord sobre la informació que ha de contenir un lexicó computacional. S'han esmerçat esforços a definir un model d'entrada lèxica de caràcter general a partir del qual es derivin els diferents lexicons, de manera que l'aplicació del model permeti comparar-los, adaptar-los i reutilitzar-los. Es considera bàsica la informació corresponent a:

- el lema o nom de l'accepció;
- la transcripció fonètica, que serà única per a totes les accepcions d'un lema;
- la categoria gramatical;
- l'estructura argumental, o esquema sintacticosemàntic bàsic que caracteritza una entrada verbal;
- les diàtesis, és a dir, els diferents esquemes sintacticosemàntics alternants que accepta l'entrada. Es tracta d'una informació fonamental per controlar el procés d'anàlisi sintàctica i evitar la sobregeneració d'arbres d'anàlisi.
- La informació semàntica, que en els lexicons computacionals s'expressa mitjançant l'associació de l'entrada a un tipus semàntic prèviament definit en una ontologia. Per exemple, a *noi* se li podria assignar el tipus semàntic [HUMÀ] i a *ordinador* [ARTEFACTE]. Els tipus depenen de l'ontologia de què es parteix. En l'apartat següent presentem EuroWorNet, una ontologia que s'utilitza per al tractament de la semàntica en els sistemes de processament del llenguatge. La informació semàntica inclou també l'explicitació de les relacions com la sinonímia, antonímia i la hiponímia/hiperonímia.
- Les equivalències de traducció amb altres llengües.

Els lexicons computacionals no solen contenir tota la informació que hem presentat. La informació bàsica correspon a la categoria i a la llista de subcategorització. A continuació, es dóna la possible representació de l'entrada *anar* en un lexicó computacional:

<anar, [[-N], [+V], [BAR 0], [SUBCAT <SP>]]>

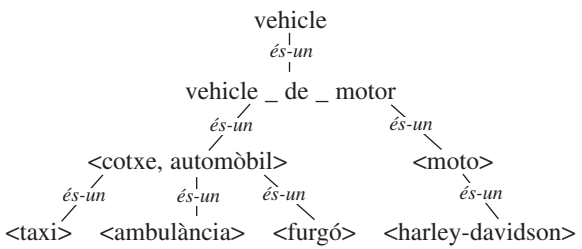
on s'indica que es tracta d'una entrada verbal ([+V]) que subcategoritza un sintagma preposicional ([SUBCAT <SP>]).

Els sistemes de representació més utilitzats per a la representació de la informació lèxica són les bases de dades de caràcter general, les bases de dades lèxiques i les bases de coneixement lèxic, que es poden basar en esquemes (*frames*) o en estructures de trets tipificades (Shieber 1986). Les representacions basades en esquemes o en estructures de trets presenten l'avantatge de poder definir relacions entre les diferents entrades del lexicó a més d'incorporar mecanismes d'herència de propietats que simplifiquen el procés de creació del lèxic i permeten aplicar-hi mecanismes d'inferència de propietats.

Cal esmentar com a representativa d'una nova concepció del lèxic, la teoria del lexicó generatiu de J. Pustejovsky (1995) que defineix un model de lexicó amb un nombre reduït de regles, principis i restriccions que donen compte d'una sèrie de fenòmens lingüístics a un alt nivell d'abstracció. El seu objectiu és constituir les bases d'una teoria del llenguatge de naturalesa semanticolèxica amb capacitat generativa.

2.3.1. Les xarxes semàntiques

Les xarxes lexicosemàntiques codifiquen relacions entre les paraules pel seu significat. Es poden considerar com un tipus de lexicó que se centra en la representació de les relacions lèxiques. Les relacions bàsiques que es tenen en compte són la sinonímia (per exemple, 'cotxe' *és un sinònim* d' 'automòbil'), la hiponímia ('cotxe' *és un* 'vehicle de motor') i la meronímia ('roda' *és una part de* 'cotxe'). Les xarxes lexicosemàntiques relacionen cada paraula, per exemple 'cotxe', amb tots els seus subtipus (o hipònims) -els tipus de cotxes que estan definits a la xarxa- amb el seu tipus semàntic o hiperònim:



Es tracta d'un recurs que s'utilitza per a millorar els sistemes de classificació documental i els sistemes d'extracció i recuperació d'informació: la possibilitat d'accedir als sinònims, als hiperònims o als hipònims d'un mot permet establir relacions entre documents que altrament no es relacionarien, i augmenta a més la capacitat d'interpretació semàntica de les preguntes dels usuaris d'aquests darrers sistemes.

EuroWordNet (Miller i Fellbaum 1991) és una xarxa lexicosemàntica multilingüe que s'està constituint com a un estàndard de la codificació semàntica, en la qual les llengües estan connectades a través d'un índex interlingua de manera que per a cada sentit es poden obtenir les seves traduccions a les altres llengües.[3]

2.4. Els analitzadors sintàctics i gramàtiques formals

Els objectius de l'anàlisi sintàctica són fonamentalment dos: agrupar les paraules en unitats de nivell superior (sintagmes i clàusules) que identifiquin els constituents principals de la oració i etiquetar aquests constituents amb les funcions corresponents. Es tracta de fer explícites les relacions sintàctiques que es troben de manera implícita.

En aquest tipus de procés cal establir una distinció entre l'analitzador o programa informàtic, la tècnica d'anàlisi i la gramàtica. La gramàtica és una especificació formal de les estructures permeses en el llenguatge; la tècnica d'anàlisi és el mètode emprat en el procés d'anàlisi d'una oració per determinar la seva estructura d'acord amb el que s'ha especificat a la gramàtica; finalment, l'analitzador és el programa que, aplicant una determinada tècnica i basant-se en les dades de la gramàtica efectua l'anàlisi sintàctica. Així, en tot programa informàtic que pretengui descriure la llengua hi trobem coneixements de dos tipus: coneixement directament lingüístic i instruccions sobre com fer servir aquest coneixement. Naturalment, el coneixement lingüístic descriu algun aspecte de l'estructura lingüística de la llengua en qüestió. En molts casos, aquest coneixement es formularà en forma de regles d'estructura de constituents com les següents:

1. SN → Det N
2. SN → Det N ADJ
3. SN → Pronom

Pel que fa al nivell de profunditat es distingeixen dos tipus d'anàlisi sintàctica, el parcial o superficial i el total. Les tècniques d'anàlisi total tenen com a objectiu la construcció d'arbres d'anàlisi que representin tota l'estructura sintàctica de l'oració

[3] A http://clic.fil.ub.es/demos/ es poden consultar les xarxes del català, castellà i anglès.

de la forma més detallada possible. Aquests sistemes presenten, però, l'inconvenient de ser molt poc precisos ja que realitzar l'anàlisi sintàctica completa requereix tant informació sintàctica com semàntica i són computacionalment molt lents. Funcionen bé per a dominis específics, ja que tant el lèxic com les estructures que cal tractar són molt inferiors en nombre.

L'anàlisi sintàctica sovint és un procés intermedi en el processament de la informació, la descripció que en resulta sol usar-se posteriorment com a input per a altres processos. És per això que, en la major part dels casos, no es precisa d'una anàlisi en profunditat, fet que ha motivat l'interès a desenvolupar sistemes d'anàlisi parcial o superficial. Aquests sistemes poden identificar constituents i fragments de constituents (en anglès, *chunks*), de manera que es redueix el nombre i tipus de relacions entre aquests elements i no es pretén l'assignació de funcions; conseqüentment, l'ambigüitat es redueix de forma important i l'anàlisi, encara que parcial, conté molt pocs errors. Així en un sintagma nominal com 'la lluita per la vida', s'identificaran dos *chunks*: 'la lluita' i 'per la vida', que corresponen a un sintagma nominal i a un sintagma preposicional. No s'indica la relació entre aquests sintagmes ni es construeix el sintagma nominal complex[4]. Si es vol processar textos no restringits és molt difícil que la gramàtica tingui previstes totes les estructures, elements i combinacions que es poden donar de manera espontània en un text, de manera que es pugui obtenir l'anàlisi completa de cada oració. En aquests casos, l'anàlisi superficial és la solució més recomanable.

L'anàlisi sintàctica pot consistir també en una anàlisi més profunda que permeti extreure i explicitar els constituents i les seves relacions (canvi d'ordre) de manera completa. En aquests casos cal disposar d'una gramàtica de la llengua d'àmplia cobertura i d'un lexicó computacional amb la informació necessària per desambiguar les estructures, identificar els arguments, les funcions, aplicar restriccions selectives sobre els arguments, etc. tal i com s'ha indicat en l'apartat 2.3. Cal disposar, a més, d'un sistema de restriccions que limiti la capacitat d'anàlisi a les estructures correctes. Actualment un analitzador d'aquestes característiques es pot construir per a dominis temàtics restringits en els quals es redueix de manera important l'ambigüitat lèxica i sintàctica i el nombre d'estructures i de fenòmens lingüístics que s'han de tractar.

Actualment s'observa també un interès creixent en el desenvolupament de bancs d'arbres sintàctics (en anglès, *treebanks*) o corpus analitzats sintàcticament (Civit i Martí, 2004). Encara que la seva elaboració pot ser automàtica o manual, només en aquest darrer cas es garanteix la qualitat del seu contingut. En aquest cas, el volum d'aquests corpus és reduït ja que el seu desenvolupament és molt costós. L'interès

[4] Per al tema de l'anàlisi superficial, vegeu Civit (2003) capítol 4 i Màrquez et al. (2001).

aplicatiu dels bancs d'arbres sintàctics és que constitueixen una font d'informació valuosa per a observar el comportament sintàctic de les llengües i poden servir com a font d'informació per a la inferència de gramàtiques de caràcter general i d'àmplia cobertura.

L'anàlisi sintàctica és encara un tema no resolt tant pel que fa als programes informàtics com als formalismes gramaticals i els models de llenguatge que s'utilitzen. Les Tecnologies de la Llengua han de fer front a l'anàlisi de textos d'ús real de la llengua i, en aquest marc es fa difícil imaginar un programa d'anàlisi que els pugui tractar de manera completa. De moment, l'anàlisi parcial s'ha revelat com una sortida eficient per a resoldre un bon nombre dels problemes que plantegen les aplicacions d'aquestes tecnologies.

3. Aplicacions basades en Tecnologies de la Llengua

Amb el nom d'aplicacions d'Enginyeria Lingüística es designen els sistemes informàtics complexos que resolen tasques intel·ligents utilitzant coneixement lingüístic. Aquestes aplicacions consten de programes (p.e. analitzadors), dades lingüístiques estructurades (com gramàtiques i lexicons) i una estructura informàtica que integra els diferents components.

Atès que el processament del llenguatge és un problema complex d'enginyeria del software i de gestió del coneixement que posa en joc unitats lingüístiques de naturalesa molt diversa, com ara són els morfemes, les paraules, les frases o els significats, i que requereix, a més, l'aplicació de processos tan diversos com ara l'anàlisi morfològica i sintàctica o la interpretació semàntica, és habitual que les aplicacions estiguin dissenyades de manera modular, on cada component s'encarrega d'un tipus de procés.

Les aplicacions d'Enginyeria Lingüística solen classificar-se en dos grans grups segons tractin textos llargs de caràcter narratiu o bé diàlegs. Les primeres presenten la dificultat d'haver de resoldre l'anàlisi de frases llargues, d'estructura molt diversa i difícilment predictible, però tenen l'avantatge de poder resoldre el problema de l'anàlisi a un nivell no excessivament profund i, normalment, només cal que arribin a un processament parcial del text.

Els sistemes dialogats s'usen en entorns de cerca d'informació i, en aquests casos, és necessària una comprensió "completa" del text, ja que altrament seria impossible accedir a les dades que es desitgen. Les frases solen ser més curtes, però no són més fàcils de tractar ja que els textos són més espontanis i, per tant, menys normatius, i sovintegen les referències anafòriques i les el.lipsis, que dificulten la seva interpretació.

L'aplicació més característica del tractament de la informació textual és la traducció automàtica, però actualment estan adquirint un gran relleu altres aplicacions relacionades amb la gestió de continguts, en especial en l'àmbit d'Internet, com són l'extracció i la recuperació d'informació i la classificació automàtica de documents.

Veurem a continuació algunes de les aplicacions més representatives de l'Enginyeria Lingüística.

3.1. La recuperació d'informació

La *recuperació d'informació* (Gonzalo y Verdejo, 2003), una activitat que fins fa poc es restringia a col·lectius de professionals molt concrets -els documentalistes-, ha passat a ser un dels problemes clau a què ha de fer front la Societat de la Informació. Internet es pot considerar com un gran banc de dades que pot devenir inservible si no se'l dota de sistemes de recuperació d'informació que satisfacin les necessitats dels usuaris potencials. El disseny de sistemes de recuperació d'informació s'ha de realitzar tenint en compte que els usuaris d'Internet ja no són un grup reduït d'especialistes, sinó qualsevol individu en qualsevol part del món. A més, les institucions, organismes i empreses que han de gestionar grans volums d'informació són també potencials usuaris d'aquest tipus d'aplicació.

El problema de la recuperació d'informació se centra en dues qüestions fonamentals: el filtrat de les dades i el multilingüisme. Com a filtrat de dades s'entén la selecció d'aquells documents que interessen l'usuari en base a la seva demanda o a uns determinats paràmetres predefinits: es tracta de recuperar els documents que interessen i evitar el "soroll", és a dir, la recuperació de documents irrellevants. Pel que fa al multilingüisme, cal disposar de recursos computacionals com els que hem presentat al segon apartat d'aquest article que permetin fer les demandes en qualsevol llengua i obtenir els documents en la llengua que l'usuari seleccioni.

De manera general, la recuperació d'informació té com a objectiu recuperar, a partir d'una col·lecció de documents, aquells que corresponen a la cerca de l'usuari. En els sistemes computacionals, la informació textual sol trobar-se emmagatzemada fonamentalment en dos tipus de formats: en forma de dades organitzades de forma tabular (bases de dades) i en forma de textos en el si de col·leccions de documents, com pot ser Internet.

Les bases de dades tenen associats llenguatges d'interrogació que permeten obtenir exactament allò que es desitja, és a dir, tenen sistemes d'accés estandarditzats. En el cas de les col·leccions de documents, el que l'usuari vol recuperar no sol ser una dada específica o un ítem lèxic, sinó més aviat una llista d'ítems, que en aquest cas

seran documents o parts de documents, que tracten d'un determinat tema en què l'usuari està interessat. Els sistemes de recuperació d'informació tracten d'obtenir per mitjans automàtics aquells textos d'una col·lecció de documents que són rellevants respecte d'una determinada pregunta de l'usuari.

Per poder accedir als documents, els sistemes de recuperació d'informació han d'afrontar la realització de les tasques següents:

- Indexar els documents a partir dels ítems d'informació (paraules, expressions, noms propis, col·locacions) que es consideren rellevants. Aquesta tasca es realitza mitjançant l'associació d'un llistat de paraules discriminadores (paraules de l'índex) als documents[5].
- Representar les consultes de l'usuari en un llenguatge intern interpretable, com a resultat d'haver analitzat la demanda[6].
- Determinar el subconjunt de documents, els índexs dels quals tenen una alta similitud amb la pregunta de l'usuari un cop ja ha estat interpretada.
- Avaluar la qualitat de la resposta.

El primer d'aquests processos, la indexació, consisteix a assignar a cada document un o més descriptors que el representen i que figuren en un índex predefinit. Quan l'usuari fa una cerca, es recuperen els documents indexats pel(s) terme(s) que conté. El procés d'indexació es pot fer manualment o de manera automàtica. Quan es tracta d'indexar grans volums de documents se sol recórrer a processos automàtics, ja que els resultats entre una i altra metodologia no són gaire diferents.

En el cas de la indexació automàtica cal determinar: a) de quina manera es confegirà l'índex sobre el que posteriorment es farà la cerca; i b) de quina manera s'associa cada document a un o més elements de l'índex. En la confecció de l'índex cal determinar quins tipus d'unitats s'admetran: paraules aïllades, lexies o col·locacions i si s'exclouran o no les paraules funcionals (pronoms, articles, preposicions, conjuncions...).

Si els elements de l'índex s'han simplificat i només hi figuren les formes canòniques dels mots (lemes), cal aplicar tècniques de processament del llenguatge en l'associació dels documents als termes de l'índex. Per a l'anglès, que té una

[5] En l'apartat corresponent a les tècniques d'extracció d'informació s'expliquen diferents tècniques d'indexar documents.

[6] Normalment, la interpretació de la consulta de l'usuari implica el seu processament morfològic sintàctic i la seva interpretació semàntica.

morfologia molt pobra, s'apliquen processos d'obtenció de l'arrel (*stemming*) consistents a eliminar els darrers caràcters dels mots que coincideixen amb un element d'una llista predeterminada de morfemes flexius. De manera ideal, la forma resultant hauria de ser un lema, encara que sovint no és així. A les llengües amb més riquesa flexiva s'apliquen processos de lematització basats en l'anàlisi morfològica del text. Alguns sistemes incorporen cert coneixement semàntic i l'índex s'enriqueix amb sinònims i hipònims que amplien la riquesa lingüística de les consultes.

La consulta de l'usuari es representa amb un llenguatge intern interpretable pel sistema. Les preguntes es poden fer a partir de la combinació de paraules clau amb operadors *booleans* (AND, OR), o bé a partir de frases en llenguatge natural. En aquest cas, solen interpretar-se igualment com a paraules clau combinades amb aquests operadors.

La selecció dels documents es pot realitzar de manera automàtica o interactuant amb l'usuari. La selecció automàtica es realitza recuperant aquells textos indexats per les paraules clau que ha expressat l'usuari en la cerca. En els sistemes interactius, es pot incorporar un procediment iteratiu de refinament de la consulta i, d'aquesta manera, s'adquireix el grau de precisió necessari per als objectius de l'usuari.

Des de l'any 1992 tenen lloc les Text Retrieval Conferences (TREC)[7], congressos de caràcter competitiu que tenen com a objectiu impulsar les tècniques de recuperació d'informació. La base dels TREC és proveir a tots els participants de dades i mètodes d'avaluació uniformes i d'una col·lecció de textos adient per a posar a prova el sistema que volen avaluar. L'aportació d'aquestes competicions és que incorporen mesures objectives per avaluar la qualitat dels sistemes de recuperació d'informació que es posen a prova.

Les línies en què es treballa actualment per millorar aquest tipus d'aplicació és incorporant coneixement lingüístic en la pregunta, especialment informació semàntica i tècniques de tractament dels elements elidits i les referències anafòriques. Una altra línia consisteix a millorar les tècniques d'extracció d'informació per a la confecció de l'índex, que es tracta en l'apartat següent.

Un pas endavant en els sistemes de recuperació d'informació el constitueixen el sistemes pregunta-resposta (Vicedo, 2003). En aquest cas el que es tracta de recuperar no és tant un document, sinó el fragment d'un o més documents que donen la resposta precisa a una pregunta de l'usuari. Si bé els sistemes de recuperació d'informació convencionals utilitzen tècniques bàsicament

[7] http://trec.nist.gov/

estadístiques, els sistemes de pregunta-resposta cada vegada més utilitzen tècniques de processament del llenguatge[8].

3.2. L'extracció d'informació

L' *extracció d' informació* té com a objectiu processar textos, documents, pàgines WEB, etc. per identificar-ne la informació rellevant, i així, disposar d'informació fàcilment accessible i utilitzable. Com a resultat del procés d'extracció es disposa, en els casos més simples, d'una llista de termes rellevants, que pot funcionar com a resum del propi document o com a índex per als sistemes de recuperació d'informació. En altres casos, s'obté una estructura d'informació organitzada on es recullen les dades més importants.

Les tècniques d'anàlisi que s'apliquen per extreure informació solen ser l'anàlisi morfològica (lematització) i l'anàlisi superficial del text, amb algun component per al tractament semàntic. Alguns d'aquests sistemes fan servir també informació sobre l'estructura del discurs per tal de distingir les parts d'un text que són rellevants de les que no ho són.

Tradicionalment, els sistemes d'extracció d'informació duen a terme el procés havent definit prèviament de manera esquemàtica el tipus d'informació que es vol obtenir. Ens trobem sempre en el marc d'aplicacions que tracten dominis lingüístics restringits que permeten predeterminar el tipus d'informació que es busca. Així, es defineixen plantilles (en anglès *template*) on es representa esquemàticament la informació que s'ha d'extreure: el procés d'extracció consisteix a omplir correctament els diferents camps de la plantilla amb la informació que s'extreu dels documents. En el procés d'extracció es necessari explicitar les relacions semàntiques existents entre els diferents elements seleccionats. Posteriorment, la informació s' incorpora a bases de dades relacionals per a la seva posterior recuperació. Aquestes plantilles poden constituir també la base per a la confecció de resums o bé l'índex per a la posterior recuperació dels documents.

Els congressos del MUC[9] (*Message Understanding Conferences*), que es realitzen cada dos anys des del 1987, han donat un impuls decisiu al desenvolupament de les

---

[8] Es poden consultar sistemes de pregunta-resposta accesibles a través d' Internet:
Omnibase, http://www.ai.mit.edu/projects/infolab/globe.html
IO search engine, http://www.ionaut.com:8400/
Webclopedia, http://www.isi.edu/natural-language/projects/webclopedia/
AskJeeves, http://www.ask.com
LCC, http://www.languagecomputer.com/
AnswerBus, http://www.answerbus.com
Q-GO, http://www.q-go.com/es/solutions/

[9] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

tècniques d'extracció d'informació. Aquests congressos estan concebuts en termes d'una competició en què els grups de recerca participants posen a prova el seu sistema. Inicialment reben un corpus d'entrenament i, en un màxim de sis mesos, han d'adaptar el seu sistema d'extracció al domini del corpus. Un cop acabat aquest període d'entrenament, es lliura als participants una nova col·lecció de textos i els han de processar amb el sistema ja adaptat. Els resultats de cada grup es comparen amb les plantilles "correctes" confegides a mà pels organitzadors. El grup guanyador és aquell que ha obtingut les plantilles més semblants a les que serveixen de test. En pocs anys s'espera tenir disponibles aplicacions d'ús general.

Un dels problemes que s'han posat en evidència gràcies als MUC és que hi ha un component important de subjectivitat a l'hora de determinar el que és rellevant en un text i, per tant, no hi ha un acord sobre el que s'ha d'extreure. La coincidència en els resultats en l'extracció manual oscil·la entre el 60% i el 80% i, per tant, es fa difícil d'avaluar la qualitat dels sistemes d'extracció automàtica. Els organitzadors dels MUC han posat un èmfasi especial a determinar mesures objectives, la cobertura i la precisió, per definir el grau de rendiment dels diferents sistemes. Presentem amb un exemple en què consisteix aquesta aplicació de manera paradigmàtica. Suposem que tenim la notícia següent:

> *El grup INCE va tenir pèrdues netes en el primer trimestre de 2001 de 49 milions d'euros, enfront d'un benefici net de 38 milions en el mateix període de 2000. Aquestes pèrdues es deuen a la seva divisió de cel·lulosa, afectada pel preu de la pasta de paper (un 17% menys)*[10].

D'aquesta notícia se n'haurien d'extreure dues plantilles, que corresponen a les dues accions bàsiques que es descriuen, tot i que en la segona el verb és el·líptic. La darrera frase no correspon a una nova acció, sinó que és una explicació complementària i no hi queda representada (Figura 3).

Resultats-1:
Situació: resultat relatiu
Organització: INCE
Signe: negatiu
Quantitat: 49 milions d'euros
Període: 01-01-2001 a 31-03-2001

Resultats-2:
Situació: resultat relatiu
Organització: INCE
Signe: positiu
Quantitat: 38 milions d'euros
Període: 01-01-2000 a 31-03-2000

Figura 3.

[10] Extret de Gonzalo i Verdejo (2003).

L'extracció d'informació té múltiples aplicacions potencials. La més immediata és l'alimentació de bases de dades relacionals que després poden ser consultades i actualitzades amb nova informació. Una altra de les aplicacions més clares dels sistemes d'extracció d'informació és la millora dels cercadors d'Internet (vegeu l'apartat anterior, *recuperació d'informació*).

Els sistemes d'extracció d'informació actuals procedeixen de la combinació de dues tècniques ben diferents: la indexació documental i els sistemes de comprensió de textos de la Intel·ligència Artificial. Podem distingir-hi dues línies metodològiques. D'una banda hi ha els sistemes d'extracció basats en mètodes empírics (estadístics) i, d'una altra, els sistemes basats en el coneixement.

Els primers es recolzen en models estadístics i en tècniques d'aprenentatge automàtic. Segons els documents que s'han de processar, aquests mètodes poden donar molt bons resultats i, a més, tenen un temps de resolució molt ràpid. L'ús de coneixement superficial per a l'extracció d'informació té l'avantatge de ser ràpid d'implementar i que s'adapta fàcilment a nous dominis. Entre els aspectes negatius cal destacar la manca de precisió en els continguts que es recuperen.

Pel que fa als sistemes basats en el coneixement, aquests requereixen la construcció prèvia de fonts de coneixement lingüístic, com són gramàtiques, diccionaris i bases de coneixement. Aquests sistemes tenen el repte de superar l'anàlisi de dominis restringits i sembla que la solució a aquest problema passa per la incorporació de coneixement lingüístic de caràcter general. Alguns models de representació del coneixement lingüístic com les LCS (*Lexical Conceptual Structures*) de Jakendoff (1988) s'estan utilitzant per a aquest fi.

Hi ha una sèrie d'aplicacions que utilitzen tècniques de processament molt properes a les de l'extracció d'informació, algunes de les quals, com la indexació i l'alimentació de bases de dades ja han estat esmentades. Cal esmentar, a més, la classificació de documents, que consisteix a assignar a cada document, d'una col·lecció, una o més categories (classes) en termes d'una classificació predefinida.

3.3. La traducció automàtica

S'entén per traducció automàtica aquella branca de la Lingüística Computacional que s'ocupa del disseny, la implementació, l'avaluació i l'ús de programes informàtics per traduir textos d'una llengua a una altra de manera automàtica. En aquest sentit, la traducció automàtica és l'àrea d'aplicació paradigmàtica de la Lingüística Computacional. Tot i així, és també una de les aplicacions més difícils de tractar, perquè traduir no és simplement substituir paraules d'una llengua a una altra, sinó que implica tot un procés molt més complex en què s'ha de tenir en compte no només la correcta elecció de les paraules en la llengua de destí sinó també cal sovint fer canvis a nivell sintàctic.

Malgrat que la investigació en aquesta àrea avança de manera molt ràpida, les dificultats inherents a la traducció automàtica són moltes. La tasca de traducció implica la resolució de problemes derivats de la diferència estructural entre les llengües, les dificultats per expressar segons quins conceptes en llengües culturalment distants, els problemes inherents de la terminologia tècnica, etc. A aquests problemes cal afegir, a més, aquells que són propis de la traducció automàtica, que deriven de la manca de models complets de comprensió automàtica de les llengües, de la inexistència de models del procés de traducció i de la falta de documentació sobre el tipus d'accions que normalment du a terme el traductor.

Però sens dubte el problema principal de la traducció automàtica és l'ambigüitat inherent del llenguatge natural, ambigüitat que el programa informàtic ha de resoldre a nivell morfològic, lexicosemàntic i sintàctic. No ens hem de sorprendre, per tant, que es tracti d'una aplicació que necessiti disposar de tots els recursos lingüístics que tinguin a l'abast, és a dir, analitzadors morfològics, lematitzadors, desambiguadors, lexicons computacionals, analitzadors sintàctics, gramàtiques, etc.

Hi ha tres factors a tenir en compte a l'hora de dissenyar un sistema de traducció automàtica:

- Si el sistema serà bilingüe o multilingüe (és a dir, si traduirà més d'una llengua).
- Si el mode d'operació serà directa o intervencionista.
- L'estratègia que seguirà, és a dir si es tractarà, bàsicament, d'un sistema de traducció directa, basat en transferència o en memòries de traducció[11].

Els sistemes de traducció directa són dels primers que apareixen al mercat i es basen en diccionaris monolingües i bilingües molt grans. Es tracta de sistemes que disposen d'un coneixement lingüístic molt limitat, reduït sovint a un petit mòdul d'anàlisi morfològica i que en cap cas realitza una anàlisi sintàctica. Són, per tant, sistemes molt simples, normalment força ràpids, perquè es limiten a la consulta dels diccionaris, però amb una qualitat de traducció molt baixa.

Els sistemes de traducció basats en transferència són sistemes molt més complexos, en els quals la traducció s'efectua en tres fases: una d'anàlisi, una de transferència i una de generació. En la fase d'anàlisi es procedeix al processament morfològic i sintàctic amb l'objectiu d'obtenir un arbre que descrigui l'estructura sintàctica de

[11] Deixem de banda els sistemes interlingua perquè actualment gairebé no s'utilitzen. Vegeu (Alonso 2003) per a més informació.

l'oració. És a partir d'aquest arbre sintàctic, que s'inicia la fase de transferència que consisteix a explicitar per a cada element de l'arbre de la llengua font, l'estructura que li correspon a la llengua destinació i a substituir els ítems lèxics de la llengua origen per la seva traducció en base a un diccionari bilingüe. El resultat final de la fase de trasferència és un arbre sintàctic on cada paraula de la llengua d'origen s'ha substituït per una paraula en la llengua de destinació i amb algunes indicacions estructurals que s'hauran de tenir en compte en la següent fase. En la fase de generació es du a terme tota una sèrie de tasques pròpies de la llengua de destinació, com per exemple la generació de les formes flexives adequades, la correcta col·locació de les paraules en funció de les regles d'ordre de constituents de la llengua de destinació, inclusió o eliminació de peces lèxiques, etc.

La qualitat dels sistemes de traducció depèn molt del tipus d'informació que s'analitzi, òbviament si entren en joc aspectes que tenen a veure amb la pragmàtica la qualitat de la traducció pot disminuir radicalment, però si en canvi depèn d'informació morfològica, lèxica, sintàctica i, en part, semàntica, es poden obtenir traduccions de qualitat. Un altre dels factors a tenir en compte és la proximitat de les llengües a traduir, els resultats de traducció són normalment millors entre llengües properes (per exemple, entre el català i el castellà o el gallec i el català) que no pas entre llengües molt diferents (com per exemple el català i el rus). De la mateixa manera, és molt important també considerar el tipus de text a traduir, un text tècnic presenta menys problemes que un text d'opinió.

Els sistemes de traducció automàtica es caracteritzen des de fa temps pel tractament de dominis lingüístics restringits o subllenguatges, per la valoració del paper del traductor –que s'ha incorporat als equips de recerca–, per l'acceptació de diferents nivells o graus d'automatització en el procés de traducció i pel desenvolupament d'entorns d'usuari que faciliten la tasca d'actualització de la informació necessària per al funcionament dels sistemes.

Encara que actualment una part de la investigació es dedica al desenvolupament de sistemes que siguin capaços de gestionar diferents dominis lingüístics, no és possible disposar de sistemes comercials de qualitat que siguin capaços de tractar parells de llengües sense cap mena de restricció i, per tant, amb cobertura universal. En aquesta línia, una de les característiques que incorporen alguns dels sistemes actuals és la transportabilitat, definida com la possibilitat de canviar el domini de l'aplicació o les llengües implicades de forma no massa complexa, de manera que l'usuari pugui definir el nou domini o incorporar les dades lingüístiques necessàries sense la intervenció d'un expert informàtic.

Un altre dels aspectes importants dels sistemes de traducció automàtica d'avui dia és la flexibilitat del seu manteniment. La majoria dels sistemes actuals disposen d'entorns interactius per tal que l'usuari –ja sigui lingüista, traductor o lexicògraf–

pugui modificar o augmentar d'una forma àgil i còmoda les dades lingüístiques (especialment les terminològiques) a fi i efecte d'anar introduint les modificacions que milloraran el funcionament del sistema.

Un cop demostrada la inviabilitat dels sistemes de traducció automàtica 'clàssics' desenvolupats en els anys setanta, que pretenien una traducció totalment automàtica i d'alta qualitat per a qualsevol parell de llengües, els objectius d'aquesta aplicació es van orientar cap al desenvolupament de recursos d'ajuda a la traducció, entre els quals es troben els lèxics i les bases de dades terminològiques multilingües. Actualment[12] són també de gran ajut les memòries de traducció, que consisteixen a emmagatzemar de manera paral·lela els textos que s'han traduït i les seves traduccions, tot i establint les correspondències oració per oració i, fins i tot, per segments inferiors al de l'oració. A la llarga, aquesta estratègia permet disposar d'un gran banc de dades de textos traduïts, de manera que davant la traducció d'un nou text el sistema busca a la base de dades quins fragments ja s'han traduït anteriorment i aplica la traducció de manera automàtica.

Enfront dels primers sistemes de traducció que estaven integrats en un entorn de hardware monousuari i es basaven en estratègies de traducció directa, els sistemes actuals funcionen en entorns PC, admeten treballar en forma multilloc de treball i permeten adpatacions personals del lexicó i de les memòries de traducció.

3.4. Les tecnologies de la parla

Les tecnologies de la parla tenen com a objectiu principal facilitar la interacció entre les persones i els sistemes informàtics de la manera més natural possible. En el que hem presentat fins ara, l'objecte de tractament informàtic han estat textos escrits. Ara bé, la manera natural de comunicar-nos és mitjançant la veu i els sistemes informàtics han de ser capaços de superar les restriccions que imposen el teclat i una pantalla. Les tecnologies necessàries per al tractament de la veu són específiques, de manera que constitueixen una disciplina diferenciada respecte del tractament de textos escrits.

El tractament de la parla implica resoldre dos problemes fonamentals: d'una banda, la producció de veu o síntesi i, d'una altra, el reconeixement de veu. El primer, la síntesi, és tècnicament més senzill, tot i que obtenir veus "naturals" és encara un

[12] A continuació s'indiquen les pàgines web on es poden consultar els diferents sistemes que hi ha de traducció automàtica del català, inclosa la variant valenciana:

Internostrum, http://www.torsimany.ua.es/

SALT, http://www.cult.gva.es/dgoiepl/salt/

AutomaticTrans, http://www.automatictrans.es

Comprendium, http://comprendium.es

problema no resolt. El segon, el reconeixement de la parla, és un tema més complex ja que el sistema informàtic ha de ser capaç de segmentar la cadena fònica de la manera correcta la qual cosa implica tractar un gran nombre de variables a més d'altres factors com poden ser el soroll, errors, etc.

La síntesi de veu consisteix a generar un senyal vocal, és a dir, parla, a partir d'un sistema informàtic que utilitza dades i regles introduïdes prèviament (Llisterri 2003). Els conversors de text en parla, constitueixen un subtipus d'aquests sistemes, i el seu resultat és l'equivalent a una lectura en veu alta d'un text. En l'actualitat es disposa de sistemes que permeten generar missatges verbals en diferents condicions, el problema és que no presenten la naturalitat pròpia de la parla humana. Les línies de recerca actuals s'orienten a millorar la naturalitat dels conversors text-veu incidint especialment en la prosòdia i a millorar la seva flexibilitat tractant que es reflecteixi l'estat emocional que hauria d'acompanyar el text. Es tracta de qüestions no menyspreables ja que, de vegades, constitueixen l'únic mitjà de comunicació de les persones que han perdut la capacitat de comunicar.

El reconeixement de la parla consisteix en la generació automàtica d'un conjunt de símbols discrets (en general, un text escrit) a partir del senyal de parla (Llisterri 2003): es tracta, per tant, de convertir el senyal fonètic en una representació que pugui tractar un sistema informàtic. Els sistemes de reconeixement es caracteritzen per tres paràmetres bàsics: el tipus d'enunciats que poden tractar, el nombre de locutors que accepta un sistema i el vocabulari que es capaç de reconèixer l'aplicació.

Els sistemes de reconeixement de la parla poden tractar tant paraules aïllades com la parla continua., el que s'anomena dictat automàtic. En el mercat ja es troben productes que permeten aquest funcionalitat, però presenten encara limitacions: solen ser sistemes que necessiten fases prèvies d'aprenentatge (l'usari ha de fornir al sistema mostres de la seva veu per tal que la pugui reconèixer), la seva qualitat depèn de si tracten dominis restringits, i no estan exempts d'un percentatge d'error.

L'àmbit d'aplicació dels sistemes de reconeixement de la parla pot ser tant domèstic com industrial, i resulta especialment útil per a persones amb limitacions de mobilitat ja que els facilita la interacció en un entorn prèviament acondicionat: amb la veu es poden encendre i obrir els llums, pujar i abaixar es persianes, etc. També es poden usar aquests sistemes per a resoldre tasques en què és habitual dictar la informació, com ara els informes mèdics, els documents legals, etc. En determinats contextos, poden substituir el que fins ara s'ha dut a terme amb una gravadora, amb l'avantatge que posteriorment no cal transcriure el text.

Les línies de recerca en reconeixement de veu s'orienten a millorar la qualitat dels resultats, i a inserir-los en el marc d'aplicacions basades en el diàleg, com poden ser les consultes telefòniques a centres d'informació, els serveis d'informació ciutadana o la traducció automàtica per telèfon, on es combina amb d'altres tecnologies, tant de la imatge com del text.

3.5. D'altres aplicacions

Fins aquí hem presentat les aplicaciones més característiques de les Tecnologies de la Llengua. Es tracta però d'una àrea de recerca i desenvolupament dinàmica que està generant noves aplicacions a mesura que cal donar resposta a les noves necessitats que planteja la pròpia dinàmica de les Indústries de la llengua. Podem citar, en aquesta línia, la classificació de documents i la producció de resums, camps d'aplicació propers a l'extracció d'informació.

En la classificació de documents l'objectiu és assignar a cada un dels documents que s'han de tractar una o més categories d'entre les que proposa una ontologia o una taxonomia. Els classificadors automàtics es basen habitualment en la presència de determinades paraules representatives o característiques que permeten associar el document a la classe.

Els sistemes de resum automàtic solen consistir sovint en una llista d'expressions simples o d'expressions multiparaula que el sistema ha considerat rellevants com a representatives del seu contingut. El problema és determinar quines són aquestes frases o expressions. Els primers sistemes feien la selecció basant-se en criteris de freqüència. Posteriorment s'han tingut en compte aspectes referents a l'estructura del document o del propi text per donar més rellevància a determinats fragments. Aquests sistemes recolzen fonamentalment en mètodes estadístics, tot i que actualment s' incorpora cada cop més coneixement lingüístic per tractar determinades formes d'ambigüitat, resoldre les anàfores i la correferència.

Els beneficis d'aquestes aplicacions són evidents: la classificació automàtica de documents, associada a un sistema de flux massiu d'informació com pot ser Internet, permet seleccionar els documents interessants segons un determinat perfil, i així tenir-los disponibles per a la seva posterior recuperació. La confecció de resums ens estalvia la tediositat d'haver de llegir molts textos quan tractem de trobar, dins d'un domini, els que tracten d'un tema específic. Com fàcilment es pot endevinar, ambdues tècniques, associades amb l'extracció d'informació, tenen un gran futur en la gestió de continguts a l'àmbit de les grans xarxes d'informació.

Bibliografia

Alonso, J.A. (2003): "La traducción automática", dins *Tecnologías del lenguaje*: 94-129 (Barcelona: Editorial UOC).

Chomsky, N. (1957): *Syntactic Structures* (The Hague: Mouton).

Chomsky, N. (1965): *Aspects of one Theory of Syntax* (Cambridge, Mass.: The MIT Press).

Civit, M. (2003): *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español*. *Monografías*, 3, de la SEPLN (Alicante: Sociedad Española para el Procesamiento del Lenguaje Natural).

Civit, M. / Martí, M. A (2004): 'Building Cast3LB: a Spanish Treebank', a *Treebanks and Linguistic Theories,* Kluwer Academic Publishers (en preparación).

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

Gonzalo, J. / Verdejo, F. (2003) "La extracción y recuperación de información" dins *Tecnologías del lenguaje*: 157-192 (Barcelona: Editorial UOC).

Harris, Z.-H. (1993): *The Linguistics Wars* (Oxford: University Press Oxford).

Jakendoff, R. (1988): *Semantics and Cognition* (Massachussets: MIT Press).

Joshi, A (1984): 'How Much Context-Sensitivity is Required to Provide Reasonable Structural Descriptions: Tree Adjoining Grammars', a Dowty, D. / Karttunen, L. / Zwicky, A. (eds.), *Natural Language Processing: Psycholinguistic, Computational and Theoretical Properties*: 190-205 (New York: Cambridge University Press).

Kaplan / Bresnan (1982): 'Lexical-Functional Grammar: A Formal System for grammatical Representation', a Bresnan (ed.), *The mental Representation of Grammatical Relations*: 173-281 (Cambridge, Mass.: MIT Press).

Kay (1985): 'Functional Unification Grammar: A Formalism for machine Translation', a *Proceedings of COLING 84*: 75-78 (California: Menlo Park).

Kittredge, R. / Lehrberger, J. (1982): *Sublanguage. Studies of language in restricted Semantic Domains* (New York: Walter de Gruyte).

Koskeniemmi, K. (1983): *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD Thesis, University of Helsinki, Department of General Linguistics.

Leech. G. (1997b): 'Introducing Corpus Annotation, a Garside / Leech / McEnery (eds.), *Corpus Annotation. Linguistic Information from Computer Text Corpora*: 1-18 (London: Logman).

Llisterri, J. (2003) 'Las tecnologías del habla', dins *Tecnologías del lenguaje*: 249-281 (Barcelona: Editorial UOC).

Martí, M.A. (1988): *Processament informàtic del Llenguatge Natural: un sistema d'anàlisi morfológica per ordinador*. Tesi Doctoral, Dept. Filología Románica, Barcelona.

Martí, M.A. (2003): "Consideraciones sobre la polisemia", dins Martí, M.A / Fernández, A. / Vázquez, G. (eds.): *Lexicografía computacional y semántica*: 61-103 (Barcelona: Edicions Universitat de Barcelona).

Martí, M. A. (coord.) (2003): *Tecnologías del lenguaje* (Barcelona: Editorial UOC).

Martí, A. / Castellón, I. (2001): *Lingüística Computacional* (Barcelona: Edicions Universitat de Barcelona).

Màrquez, Ll / Padró, Ll / i Rodríguez, H. (2001): 'Mètodes robustos en l'anàlisi del llenguatge. El processament de text no restringit, *Lingüística Computacional*: 1-68 (Barcelona: Edicions de la Universitat Oberta de Catalunya).

McEnery, G. / Wilson, A. (1996): *Corpus Linguistics* (Edinburgh: Edinburgh University).

Miller, G. A. / Fellbaum, Ch. (1991): "Semantic networks of Englis", *Cognition*, 41: 197 -229.

Moreno Sandoval, A. (2001): *Gramáticas de unificación y rasgos*. Col. Lingüística y Conocimiento. 32 (Madrid: A. Machado Libros, S.A.).

Oakley / Owen (1990): *Alvey, Britain's Strategic Computing Initiative* (Cambridge, Mass: MIT Press).

Ooi, V. B. Y. (1998): *Computer Corpus Lexicography* (Edinburgh: Edinburgh University press).

Pollard, K. / Sag, I. (1987): *Information-Based Syntax and Semantics*. Volume 1: *Fundamentals*. CSLI, Lecture Notes, 13. (California: Stanford).

Pollard, K. / Sag, I. (1993): *Head-Driven Phrase Structure Grammar*. CSLI (Chicago / London: The University of Chicago Press).

Pustejovsky, J. (1995): *The generative lexicon* (Cambridge MA: The MIT Press).

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

Sinclair, (1987): *Looking Up* (London: Collins ELT).

Taulé, M. / Martí, M. A. (2001): 'Formalismes gramaticals', *Lingüística Computacional*: 1-99 (Barcelona: Edicions de la Universitat Oberta de Catalunya).

Vicedo, J. L. (2003): *Recuperación de información de alta precisión: los sistemas de búsqueda de respuestas*. *Monografías*, 2, de la SEPLN (Alicante: Sociedad Española para el Procesamiento del Lenguaje Natural).