Ma. Antònia Martí i Antonín

Tesi Doctoral

"Processament informàtic del llenguatge natural: un sistema d'anàlisi morfològica per ordinador"

Departament de Filologia Romànica

Facultat de Filologia de la Universitat de Barcelona

Barcelona 1988

Directora:

Dra. Teresa Cabré i Castellví

Tutor:

Dr. Jesús Tusón

INDEX

I.- LA LINGÜÍSTICA COMPUTACIONAL

La Lingüística Computacional és una nova disciplina que ha sorgit de la col.laboració entre la linguistica i la informàtica. Figura com a component en àrees d'investigació diverses com l'estadística, la psicologia i la mateixa linguistica.

Les diferents aplicacions de la informàtica al llenguatge requereixen un tractament de la llengua que pot anar del simple recompte de formes fins a una certa 'comprensió' del text.

Entre les diferents aplicacions i en ordre de complexitat creixent, tenen especial interès:

1)- els programes estadístics per determinar freqüències de formes en un text;

2)- la lematització, consistent en la reducció de les formes flexives d'un text al seu lema, la forma canònica;

3)- els primers sistemes de traducció automatica basats en una correspondència paraula-paraula;

4)- els sistemes que requereixen un nivell més o

menys complex de 'comprensió' del text, com són els sistemes de traducció automàtica, els sistemes de tractament de la informació textual i les interfícies home-màquina. Normalment aquests sistemes necessiten diferents tipus de tractament lingüístic del text, com ara l'anàlisi morfològica, l'anàlisi sintàctica i l'anàlisi semàntica, els quals han donat lloc a diverses àrees d'investigació en Lingüística Computacional.

L'anàlisi morfològica automatitzada ha estat una àrea relativament poc explorada en Lingüística Computacional perquè l'anglès, llengua d'aplicació de la majoria d'investigacions, és pobre en aquest sentit. Contràriament, l'anàlisi sintàctica ha estat, des del començament, l'àrea sobre la qual més s'ha treballat en Lingüística Computacional, perquè els treballs de Harris i Chomsky havien demostrat la possibilitat de la seva formalització rigorosa i perquè es considerava que era l'aspecte més important per a la construcció d'una teoria del llenguatge i conseqüentment, per a la seva representació formal.

A mesura que les investigacions en Lingüística Computacional han anat exigint un nivell de resultats més proper al que seria un comportament 'intel.ligent', ha calgut introduir altres tipus de de coneixement com ara el semàntic i el pragmàtic, els

quals han permès millorar els resultats, però que han exigit l'establiment de certes limitacions: la introducció de la representació semàntica i, molt més encara, del component pràgmatic (1), ha demostrat la necessitat de definir per a cada estudi en què el llenguatge juga un paper important el domini sobre el qual s'aplica.

La necessitat d'incorporar aquests darrers components fa de la Lingüística Computacional una disciplina diferenciada del estudis purament lingüístics pel fet que tracta amb especial atenció la interacció entre el coneixement lingüístic i el no lingüístic, en el sentit de veure com els actes lingüístics encaixen en un context més ampli d'acció i coneixement. D'aquí ve l'interès per la descripció del context per tal de controlar la seva acció sobre el llenguatge.

És en aquest marc de recerca que la Lingüística Computacional coincideix amb una àrea d'investigació, la Intel.ligència Artificial (I.A.), que desenvolupa mitjans computacionals per reproduir el comportament intel.ligent humà. Les diferents àrees d'investigació en què es treballa en I.A. són, entre d'altres: la

---

(1) El component pragmàtic tracta la representació formal del context extralingüístic en què es produeixen els actes de comunicació.

robòtica, la representació del coneixement, la visió, la comprensió del llenguatge, la traducció automàtica, els sistemes experts, l'aprenentatge, el raonament,...etc.

Des de la perspectiva de la Intel·ligència Artificial l'estudi del llenguatge té dos objectius:

a)- facilitar la comunicació amb l'ordinador per tal que hi puguin accedir usuaris no especialistes;

b)- modelar els processos cognoscitius que entren en joc en la comprensió del llenguatge per al disseny de sistemes que realitzin tasques lingüístiques complexes com ara la traducció i el resum de textos.

La comprensió del llenguatge natural recolza en la possibilitat de transformar les dades lingüístiques en elements de coneixement representats de forma tractable per un sistema intel·ligent. Es tracta de transformar 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.

A continuació presentem les principals àrees d'aplicació de la Lingüística Computacional, la Traducció Automàtica, les Interfícies en llenguatge natural i el tractament de la informació textual, des de la perspectiva del seu desenvolupament.

I.1.- Aplicacions de la Lingüística Computacional

Les aplicacions informàtiques que tracten el llenguatge natural tenen com a primera matèria el llenguatge. Aquest fet, encara que sigui molt evident, sovint no s'ha valorat en profunditat i ha representat un problema important que ha afectat el seu desenvolupament. Durant molt de temps s'ha tendit a subvalorar els aspectes lingüístics del problema de manera que l'aplicació de la Linguística Computacional s'ha considerat com un problema d'enginyeria del software, tot reduint el llenguatge natural a un llenguatge formal més, encara que regit per una gramàtica més complexa. Es nota una tendència a extrapolar a escala real resultats que són certs només en entorns molt limitats, com si el canvi d'escala només afectés aspectes quantitatius marginals, quan en realitat afecta aspectes fonamentals que poden invalidar tota la filosofia d'un sistema.

Una aplicació del llenguatge natural suposa un maneig d'informació lingüística considerable: diccionaris, gramàtiques, reglas, relacions textuals o de diàleg, etc. Suposa també un problema complex d'enginyeria del software, així com el desenvolupament de tècniques i mètodes específics per gestionar tota aquesta informació.

L'acceptació d'aquests fets ha comportat, d'una banda, el desenvolupament de la Lingüística Computacional com a pas necessari per a la construcció de qualsevol sistema informàtic basat en el llenguatge natural; de l'altra, l'admissió del fet que el llenguatge natural, amb tota la seva complexitat i riquesa, no podrà ser controlat en un termini previsible i que el desenvolupament de les seves aplicacions fa necessari restringir-lo artificialment. A partir d' aquesta darrera evidència s'han creat els conceptes "abast conceptual" i "abast lingüístic" per definir el poder expressiu que presenta un subconjunt del llenguatge natural.

L' **abast conceptual** d'un sistema l'imposa l'especificació del propi sistema i està limitat tant pel domini semàntic, és a dir, el subconjunt del camp semàntic on es situen els conceptes implicats, com per les operacions que s'han de realitzar amb aquests conceptes. El terme 'abast conceptual' es refereix al conjunt de conceptes que ha de tractar el sistema i a les operacions que realitza amb ells.

L' **abast lingüístic** mesura la riquesa de les expressions que el sistema és capaç de controlar. En aquest cas és fonamental la riquesa morfològica i sintàctica del sistema, la qual pot anar des de l'acceptació de formes molt restrictives fins a una

utilització gairebé lliure del llenguatge.

El concepte de **subllenguatge** intenta establir aquestes restriccions basant-se en el propi llenguatge i no en les seves aplicacions. La idea d'un subllenguatge com a part del llenguatge natural, amb una gramàtica pròpia, ha estat desenvolupada per Zellig Harris en el seu treball sobre transformacions i anàlisi del discurs Mathematical Structures of Language" (Harris, 1971) :

"Determinats subconjunts de frases d'una llengua poden ser tancats per a algunes (o totes) operacions definides en el llenguatge i, d'aquesta manera, constituir un subllenguatge d'aquest llenguatge (...). Aquesta situació es manifesta de manera característica en el llenguatge de les diverses ciències, p. e.: en els conjunts de frases que descriuen sectors particulars de fenòmens estructurats. En aquest cas totes les frases satisfan certes restriccions gramaticals que no són vàlides per al conjunt de la llengua" (170).

Harris conclou que "la gramàtica d'un subllenguatge conté regles que el llenguatge viola i la gramàtica de la llengua té regles que no concerneixen per a res al subllenguatge. Així, mentre que les frases dels llenguatges científics estan incloses en el llenguatge en la seva totalitat, la gramàtica d'aquests subconjunts presenta una intersecció no buida respecte a la gramàtica del llenguatge total" (pàg.171).

Per il.lustrar aquest tema és especialment interessant l'article de Lehrberger on descriu les característiques d'un subllenguatge particular, el

dels informes meteorològics del projecte de traducció automàtica (T.A.) TAUM-METEO, i dóna la relació dels factors que permeten la caracterització d'un sublenguatge (1):

a)- una àrea temàtica limitada;

b)- les restriccions lèxiques, sintàctiques i semàntiques;

c)- les regles específiques de la gramàtica;

d)- l'alta freqüència de determinades construccions;

e)- l'estructura del text i

f)- l'ús de símbols especials.

La construcció d'una gramàtica d'un sublenguatge determinat no planteja tan sols la resolució d'un exercici merament lingüístic, sinó que exigeix una classificació dels termes rellevants i de les relacions d'un determinat tema i una representació de les seves estructures de fets més generals (2).

Presento a continuació tres àrees típiques d'aplicació de la Lingüística Computacional: els diàlegs home-màquina per accedir a diversos tipus d'informació, la traducció automàtica i diverses maneres de tractar la informació textual.

---

(1) J. Lehrberger "Automatic Translation and the Concept of Sublanguage", 1982.

(2) Z.H.Harris "Discourse and Sublanguage", 1982.

I.1.1.- Sistemes dialogats

En aquests moments les aplicacions basades en diàlegs home-màquina constitueixen el principal camp d'aplicació de la Lingüística Computacional. En primer lloc perquè el nombre d'usuaris finals que tenen contacte amb l'ordinador ha augmentat sensiblement i, en conseqüència, ha aparegut la necessitat d'apropar el llenguatge de comunicació home/màquina al llenguatge utilitzat per aquests usuaris potencials. Segonament, perquè per a un sistema de tractament del llenguatge natural representa un gran avantatge comunicar-se amb l'usuari per tal de demanar-li més informació sobre una decisió insegura, sol·licitar informació addicional, etc.

Els primers intents de construir interfícies home/màquina en llenguatge natural anaven orientats a l'accés a Bases de Dades (1), tant pel pes relatiu d'aquest tipus d'aplicacions com per la facilitat amb què certs llenguatges d'accés podien generalitzar-se fins a esdevenir un subconjunt del llenguatge natural.

---

(1) A partir d'ara utilitzarem B.D. per referir-nos a les Bases de Dades.

G. Hendrix classifica les interfícies d'accés en llenguatge natural a B.D. en tres nivells segons el tipus d'informació de què disposa el sistema (1):

- Els sistemes de nivell 1 incorporen una teoria del domini de l'aplicació molt limitada o nul.la. La traducció des del llenguatge natural al llenguatge de consulta a la B.D. acostuma a ser directa. No hi ha una representació interna de la pregunta que permeti fer-hi inferències.

- Els sistemes de nivell 2 incorporen una teoria explícita del domini de l'aplicació, és a dir que utilitzen representacions internes d'alguns objectes del domini. Solen traduir la pregunta a una forma lògica intermèdia. Normalment treballen amb models del discurs per abordar el problema de la referència.

- Els sistemes de nivell 3 incorporen tot el que he dit i a més una representació explícita de l'usuari, els objectius que es volen aconseguir, informació sobre l'usuari i el domini, plans, etc.

---

(1) G. Hendrix "Natural Language Interface", 1982

Aquesta classificació planteja un límit de complexitat creixent per a cada un dels nivells. La majoria dels sistemes que actualment estan en funcionament s'haurien de situar dins del primer nivell. Els sistemes més moderns van incorporant nous elements com:

a)- l'accés a Bases de Dades cada cop més grans;

b)- l'ús generalitzat de la pronominalització, l'elipsi, l'anàfora, etc.;

c)- formes complexes de quantificació;

d)- la correcció d'errors ortogràfics;

e)- l'extensió de l'abast conceptual;

f)- metapreguntes, és a dir, preguntes referides no al contingut de la B.D., sinó al propi sistema i a les seves capacitats.

Un dels primers sistemes dialogats que es van construir va ser l'SHRDLU de Winograd (1).

L'SHRDLU és un robot dotat d'un programa que li permet interaccionar en llenguatge natural amb una persona, executar ordres i respondre a preguntes sobre el seu entorn i les seves pròpies accions. El robot consisteix en un braç mòbil capaç d'agafar i desplaçar

---

(1) T. Winograd "Five lectures on Artificial Intelligence. Lecture 2:SHRDLU, a System for Dialog", 1977.

objectes geomètrics situats sobre una taula. Els objectes es caracteritzen per la forma, el color, la situació i la dimensió.

L'SHRLDU és un sistema integrat que combina sintaxi, semàntica i raonament, aquest entès com la capacitat de fer deduccions i connectar els fets dins d'un domini temàtic.

L'analitzador sintàctic treballa amb una gramàtica general de l'anglès i una sèrie de rutines semàntiques que contenen el tipus de coneixement necessari per poder interpretar els significats de les paraules i de les estructures. El sistema cognitiu explora les consequències dels fets, planifica com dur a terme les ordres i respon a les preguntes. Uns programes generen les respostes apropiades en anglès.

Amb l'SHRDLU es pot dur a terme una conversa del tipus:

- "Aixeca una figura vermella gran"
- "D'acord"
- "Agafa la piràmide"
- "No sé a quina piràmide et refereixes"
- "Busca la figura que és més gran que la que tu tens i posa-la dins de la caixa"
- etc.

Després d'aquesta experiència de Winograd es va imposar la idea que un sistema que hagués de comprendre efectivament el llenguatge havia de tenir un component semàntic i una certa capacitat de raonament. L' SHRDLU pot recórrer a certs coneixements semàntics del tipus: "un cub és un ésser inanimat", "només un ésser animat pot 'prendre'", de manera que pot resoldre enunciats del tipus:

-"El cub pot prendre la piràmide?"

que generen una resposta negativa. L'SHRLDU coneix també en tot moment la situació dels cubs i de les piràmides i pot raonar sobre aquest estat de fets.

Encara que el sistema operava sobre un univers molt restringit d'objectes, Winograd té el mèrit d'haver demostrat molt clarament la **necessitat de tractar simultàniament el component sintàctic i semàntic en un sistema de comprensió automàtica del llenguatge.**

El LUNAR de Woods és també un sistema dialogat d'aquesta primera època (1). Woods va crear les ATN (2) per a un programa de tractament del llenguatge natural que havia de permetre que els geòlegs del programa espacial APOL.LO interroguessin una B.D. relativa a mostres del terra de la lluna: el LUNAR va comprendre perfectament un 90 per cent de 110 preguntes que podien tenir un nivell de complexitat considerable, del tipus:

> "- Dóna'm referències sobre els basalts abissals"

o bé:

> "- Quines mostres contenen cromita?"

---

(1) W.A.Woods "Lunar Rocks in Natural English: Explorations in Natural Language Question Answering", 1977.

(2) Una TN (Transition Network) és un conjunt d'**estats** connectats per **arcs**. Els estats es poden definir com a **estats inicials** i com a **estats terminals**. Una gramàtica de xarxes de transició simple consisteix en una xarxa els arcs de la qual estan etiquetats amb paraules o categories lèxiques. Una gramàtica de xarxes de transició recursives (RTN, Recursive Transition Network) conté una xarxa per a cada nus no terminal de la gramàtica. Els arcs estan etiquetats amb símbols terminals i no terminals. Una gramàtica de xarxes de transició augmentades (ATN, Augmented Transition Network) té com a tret característic l' addició de condicions i accions associades als arcs de la xarxa. Les condicions restringeixen les circumstàncies sota les quals s'ha d'optar per un arc determinat, mentre que les accions realitzen les operacions de marcar els trets i construir les estructures.

Woods va demostrar que es podien obtenir resultats satisfactoris si s'acceptava limitar la competència d'un sistema a la comprensió d'un univers restringit. La importància de LUNAR cal també situar-la en el fet que es tracta d'una eina robusta (1) comprovada sobre un corpus real.

Cap a la meitat dels anys 70 s'incorporen progressivament als sistemes més elements propis dels nivells dos i tres. Un equip dirigit per la Dra. B. Grosz ha desenvolupat en aquests últims anys al Scientific Research Institute (S.R.I.) un sistema d'accés a bases de dades, el TEAM (Transportable English Database Acces Medium) (2), que consta de tres elements principals: un mòdul d'adquisició, un processador del llenguatge natural, DIALOGIC (3), i un mòdul d'accés a les dades. El funcionament del sistema s'inicia amb la traducció de la pregunta a una forma lògica interna que es va completant i que, finalment, es tradueix al llenguatge específic de la B.D.

---

(1) Utilitzem l'adjectiu 'robust' per indicar que es tracta d'un sistema resistent a una manipulació incorrecta per part de l'usuari. Hem emprat la traducció literal del terme anglès 'robust' ja que, de moment, en català no s'ha determinat un terme per a aquest concepte.

(2) B. Grosz "TEAM: An Experiment in the Design of Transportable Natural-Language Interfaces", 1987.

(3) B.Grosz i d'altres "DIALOGIC: A NL processing System", 1982.

El nucli del sistema, DIALOGIC, consta de cinc components:

1)- Una gramàtica general de l'anglès, DIAGRAM (1), una gramàtica d'estructura de frase augmentada.

2)- Un conjunt de traductors semàntics associats a cada regla de DIAGRAM que interpreten la funció que cada constituent compleix a la frase.

3)- Un conjunt de funcions semàntiques bàsiques que afegeixen als nusos de l'arbre d'anàlisi parts de la forma lògica que deriva dels constituents del nus.

4)- Un mecanisme de determinació de l'abast dels quantificadors i la construcció de la forma lògica de l'expressió a partir dels fragments i

5)- Un conjunt de funcions pragmàtiques bàsiques que resolen ambiguitats i indeterminacions en la forma lògica (els referents dels pronoms, l'àmbit dels sintagmes preposicionals en les frases nominals complexes, aposicions, etc.).

El mòdul d'accés a les dades tradueix, primer, la forma lògica generada pel DIALOGIC al llenguatge de consulta de la B.D., després s'accedeix a la informació i, seguidament, es genera la resposta.

---

(1) J. Robinson "Diagram: A Grammar for Dialogues", 1982 .

Per dur a terme totes aquestes tasques, TEAM ha de disposar de:

a)- Informació dependent del domini:

-informació lèxica, sintàctica i semàntica de les paraules;

-informació sobre conceptes: esquema conceptual, relacions taxonòmiques, definició de les propietats que pot tenir un objecte, tipus d'objectes que poden ser arguments d'un predicat, etc...

b)- Informació independent del domini que DIALOGIC utilitza per traduir les preguntes a formes lògiques que prèviament s'han formulat: regles sintàctiques, restriccions sintàctiques, regles semàntiques i pragmàtiques, regles de recerca de referents, regles de determinació de l'àmbit dels quantificadors, etc.

Durant els últims anys, els investigadors d' I.A., influits per l'èxit de les interfícies d'accés a B.D. en llenguatge natural, han definit altres camps potencials d'aplicació d' aquests sistemes com són: l'actualització de B.D., l'accés a sistemes experts i l'accés a sistemes operatius i a sistemes tutors.

I.1.2.- La traducció automàtica

La traducció automàtica (1) representa un camp arquetípic de les aplicacions de la Lingüística Computacional, en primer lloc perquè històricament va ser el primer intent de processament del llenguatge natural i, segonament, perquè el fet d'adquirir un text en una llengua i tornar-lo traduit a una altra dóna la idea, gairebé sempre falsa, que el text ha estat comprès i aquest és, en el fons, l'objectiu de la I.A.

D'una manera esquemàtica, el procés de traducció consisteix en el pas d'un text expressat en una llengua font (llengua de la qual es tradueix, 'source language') a un text expressat en una llengua objectiu (llengua a la qual es tradueix, 'target language').

La traducció es desenrotlla en dues fases: anàlisi o comprensió del text font ('source text') i síntesi o generació del text objectiu ('target text').

Les tècniques lingüístiques utilitzades en T.A. cal considerar-les des de dues perspectives: la profunditat de l'anàlisi i les fonts de coneixement lingüístic.

---

(1) A partir d'ara, T.A.

En el disseny d'un sistema de T.A. s'ha de definir el grau de profunditat o de comprensió que es vol assolir amb l'anàlisi (1). El grau de profunditat del mòdul d'anàlisi determinarà el nivell de profunditat a partir del qual s'iniciarà la síntesi i el disseny del mòdul de síntesi en general.

Els sistemes més senzills de T.A. consisteixen en la traducció directa des del llenguatge font al llenguatge objectiu. Aquests sistemes, en els quals l'anàlisi es redueix al mínim, tenen l'inconvenient que els diccionaris es construeixen en funció de la llengua a la qual es vol traduir i per tant no es poden utilitzar per traduir a una altra llengua: s'han de refer totalment. Entre els sistemes basats en aquesta estratègia cal destacar: el GAT, el SYSTRAN i el PAHO (SPANAM i ENGSPAN).

A l'extrem oposat de la T.A. directa es troben els sistemes que construeixen una representació profunda independent del llenguatge font i del llenguatge objectiu. Es tracta dels sistemes "interlingua". A partir d'aquesta representació s'ha de poder traduir a qualsevol llengua. De moment, però, encara no s'ha

---

(1) J. Slocum "Machine Translation: Practical Issues", 1987.

A. B. Tuckner "Current strategies in machine translation research and developement", 1987.

pogut especificar com ha de ser aquest tipus de representació excepte per a dominis lingüístics molt restringits i cenyits a l'aspecte semàntic.

En els sistemes basats en mecanismes de transferència, el text font és analitzat en una estructura més profunda que l'exigida per a la traducció directa, però no tan independent del llenguatge font com la "interlingua". Es tracta d'un estadi intermedi de representació entre els sistemes de T.A. directa i els sistemes "interlingua".

El procés de transferència transforma l'estructura construida pel mòdul d'anàlisi corresponent al text de la llengua font en l'estructura equivalent de la llengua objectiu. A partir de l'estructura transferida es genera la traducció a aquesta llengua (Fig.1). Per a N llengües, aquest mecanisme requereix N mòduls d'anàlisi, N mòduls de síntesi, N*(N-1) mòduls de transferència i N*(N-1)/2 diccionaris bilingües. Aquest és el model que utilitzen la majoria dels sistemes que s'han desenvolupat i es desenvolupen els darrers anys: l'ATLAS-I i II, el METAL, el GETA, l'EUROTRA, etc.

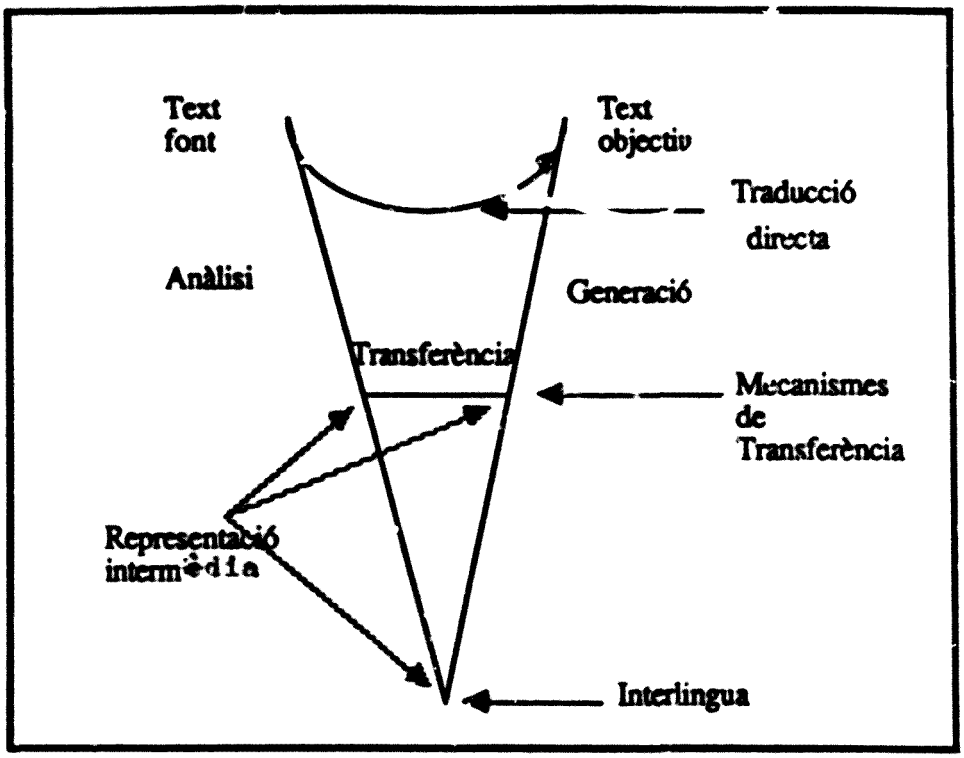


Fig.1

Una altra perspectiva des de la qual es poden considerar els sistemes de T.A. és la del tipus de coneixement lingüístic sobre el qual recolzen. Els sistemes més senzills, com el SYSTRAN i el GAT, basen tota l'estratègia de la traducció en el lexicó. Les regles d'aquests sistemes són molt específiques i la seva transportabilitat, nul.la. Es tracta dels sistemes de traducció directa que ja hem esmentat.

La font de coneixement més habitual en els sistemes de T.A. és la sintaxi, encara que s'utilitzen també altres fonts de coneixement que permeten reduir el

nombre d' interpretacions com en el cas dels sistemes de T.A. METAL, ATLAS-I, EUROTRA, GETA, etc.

La tercera font de coneixement correspon a la semàntica. Tot i que actualment no existeix cap formalisme general de regles semàntiques convincent ni àmpliament acceptat, alguns sistemes com l' EUROTRA i l'ATLAS-II utilitzen parcialment formalismes de tipus semàntic: en el cas de l'EUROTRA s'utilitzen esquemes de recció ('frames') per representar l'estructura de casos dels verbs i l'ATLAS-II utilitza xarxes semàntiques per a la representació de l'estructura de transferència.

Segons el grau d'intervenció humana es poden classificar els sistemes de traducció automàtica en tres categories (1):

A)- Els sistemes de T.A. que pretenen realitzar la traducció sense intervenció humana. No tenen ni pre-processament ni post-edició. Un sistema d'aquest tipus és absolutament responsable de tot el procés de traducció des del text font fins al text traduit i representen el grau més alt d'ambició en l'escala de la traducció automàtica.

---

(1) J. Slocum "Machine translation: Practical Issues", 1987.

B)- Els sistemes de T.A. assistida poden ser o bé sistemes de traducció automàtica que requereixen l'ajut humà, o bé sistemes d' ajuda automàtica a la traducció humana. Els primers són responsables de la traducció, però al llarg del procés sol.liciten l'ajuda humana bé per desambiguar algun mot o per saber on cal inserir un determinat arbre d'anàlisi sintàctica, bé per elegir una determinada traducció d' una paraula o frase entre una sèrie de candidates, etc. En els sistemes d'ajuda automàtica a la traducció és l'home qui s'encarrega de la traducció però interacciona amb el sistema per fer consultes al diccionari o a un thesaurus, per accedir a bancs de dades terminològiques, per obtenir exemples d'ús d'una paraula o frase, etc.

Alan Melby (1) distingeix tres nivells d'interacció home-màquina en els sistemes de T.A. El primer, el menys sofisticat, pressuposa un entorn de processament de textos amb mitjans per accedir a diccionaris en línia i a enciclopèdies. El segon nivell afegeix sistemes de correcció d'errors ortogràfics, concordances i diccionaris textuals. El tercer nivell inclou un cert grau de processament automàtic.

---

(1) A. Melby "On human machine interaction in translation", 1987.

Com a exemples de sistemes de traducció assistida per ordinador tenim el LOGOS, el Weinder i l'ALPS.

C)- Els bancs de dades terminològiques (TD, Terminology Data banks) representen el nivell més baix d'automatització en els sistemes de T.A., ja que s'hi accedeix no durant el procés de traducció, sinó abans que es produeixi. En alguns casos s'hi pot accedir interactivament durant la traducció, d'altres vegades només s'hi pot accedir mitjançant llistats d'àrees determinades de terminologia.

A continuació presentem una breu referència, des d'una perspectiva històrica, de diferents sistemes de T.A. interessants per les innovacions que han incorporat, per haver estat la base dels sistemes actuals o, finalment, per tractar-se de sistemes en funcionament o en plena actualitat (1).

Els primers intents de construcció de traductors automàtics daten dels anys 50. Es partia d'idees molt utòpiques ja que es creia que el problema de la traducció es podria resoldre amb un cost de temps raonable, sense limitacions en el domini i en un lapse de temps curt. Com que els resultats es feien esperar

---

(1) J. Slocum "A Survey on Machine translation: its History, Current status and Future Prospects", 1985

el govern nord-americà va nomenar un comitè perquè avalués les investigacions sobre traducció automàtica. Les conclusions d'aquest comitè, conegudes amb el nom d'informe ALPAC, es van fer públiques el 1964 i concloïen que la traducció automàtica no tenia cap futur immediat ni tècnic ni econòmic. Arran d'aquest fet es van suspendre totes les subvencions a aquesta àrea de recerca tant als EEUU com a Europa i al Japó.

Pertanyen a aquesta primera època alguns sistemes interessants tant des del punt de vista històric com perquè constitueixen la base dels sistemes actuals.

El primer de tots va ser el GAT (Georgetown Automatic Translation), que es va iniciar el 1952 amb el suport del govern dels Estats Units.

El 1961 es crea el sistema CETA al Centre d'Etudes pour la Traduction Automatique de Grénoble, una de les institucions que més ha contribuit al desenvolupament de les aplicacions de llenguatge natural a Europa. El sistema es proposava la traducció de textos del rus al francès La base teòrica de CETA era la de "interlingua" i gràcies a aquest projecte es va palesar la dificultat de desenvolupament d'una estratègia d'aquest tipus. El sistema es va abandonar el 1972 i a continuació s'inicià el projecte GETA, basat en mecanismes de transferència. És un sistema actualment vigent.

El 1961 la Universitat de Texas va fundar, amb ajuda del govern, el Linguistics Research Center (L.R.C.) on es va desenvolupar el projecte METAL (Mechanical Translation and Analysis of Languages) per traduir textos de l'alemany a l'anglès. L'any 1980 el va comprar l'empresa alemanya Siemens i s'hi van introduir algunes modificacions. Actualment s'està treballant en la síntesi al castellà a partir de l'alemany.

El 1965 la Universitat de Montreal amb ajuda del govern canadenc va subvencionar un projecte de traducció automàtica , el TAUM (Traduction Automatique de l' Université de Montréal). Va ser el primer projecte de traducció automàtica dissenyat a partir de la idea de transferència. El TAUM-METEO (1), dedicat a la traducció de butlletins meteorològics de l'anglès al francès, és potser l'únic dels sistemes actuals que es pot considerar totalment automàtic i de traducció d'alta qualitat. Això és així perquè tracta un domini molt restringit tant pel que fa al vocabulari com a les estructures sintàctiques

---

(1) El professor Sergei Nirenburg, del Center for Machine Translation de Carnegie-Mellon, va manifestar en la taula rodona sobre T.A. celebrada durant la III Reunió Anual de la SEPLN (2-3 juliol 1987) que de fet és l'únic sistema de T.A. efectiu que existeix.

que ha de processar.

Les causes del fracàs dels primers sistemes van ser el seu abast massa ambiciós i la inadequació de les eines i tècniques utilitzades. Els sistemes que es desenvolupen a partir dels anys 70 tenen en compte tots aquests factors i presenten una sèrie de característiques comunes:

- El sistema ha de respondre a un equilibri entre objectius i mitjans.

- S'aillen en les dues llengües (font i objectiu) els sublleguatges adequats, si és que existeixen.

- S'estudien els requeriments de qualitat de la traducció, que es podrà millorar augmentant la complexitat de les eines, lingüístiques i informàtiques, o augmentant la intervenció humana.

- En tots aquests sistemes és comuna la separació clara dels algorismes, per una banda, i de les gramàtiques i diccionaris, per una altra.

- Tots els sistemes admeten la necessitat d'una o diverses teories lingüístiques sobre les quals basar-se. No es posa en dubte la interacció dels nivells

morfològic, lèxic, sintàctic, semàntic i textual en diversos graus i maneres.

Actualment gairebé tots els sistemes estan dirigits per la sintaxi; la semàntica sol incloure's en forma de restriccions sobre les regles sintàctiques per després jugar un paper en fases de desambiguació, transferència, etc. En tots els sistemes sol ser necessària una fase prèvia d'anàlisi morfològica, tema que es tractarà en el proper capítol.

Entre els sistemes actuals cal destacar METAL (en la seva nova versió), ATLAS-I i ATLAS-II, els sistemes que s'estan desenvolupant en el marc del projecte japonès de la cinquena generació i EUROTRA, que han començat a desenvolupar a partir de 1982 els països de la CEE.

I.1.3.- Sistemes de tractament de la informació textual

Aquesta àrea d'aplicació de la informàtica al llenguatge tracta tant de l'ajuda a l'edició de textos com de l'adquisició, emmagatzemament i accés a la informació textual.

I.1.3.1.- Edició i correcció de textos

Les dues àrees fonamentals sobre les quals es treballa en l'edició i correcció de textos són la partició sil.làbica i la correcció d'errors (1).

La partició de les paraules pel lloc correcte quan no caben senceres en una línia és un dels problemes bàsics que es plantegen en l'edició i correcció de textos. Es tracta d'una problemàtica que admet un ampli ventall de solucions pel que fa al grau de complexitat i eficàcia. Una primera solució, la més simple, pot ser no permetre les particions. Un grau major de complexitat resolutiva consisteix a disposar d'un diccionari expandit amb indicadors, per a cada forma, dels llocs on es poden permetre les segmentacions. La solució més acurada hauria

---

(1) J.L. Peterson Computer programs for spelling correction, 1980

d'incorporar un sistema d'anàlisi morfològica per al tractament de la flexió i evitar així l'emmagatzemament en el diccionari de les formes flexives conjuntament amb regles de descomposició sil·làbica.

Els sistemes més senzills de correcció d'errors ortogràfics consisteixen en programes que proporcionen una llista de totes les formes del text. Una persona ha de comprovar posteriorment si a la llista hi ha errors. En aquests casos la llista està ordenada per ordre alfabètic o bé per freqüències d'aparició dels mots.

Un altre sistema consisteix en la utilització d'una llista de formes correctes (un diccionari). Primer s'elabora una llista de totes les formes del text i es comprova si són al diccionari de formes correctes. Si es troben al diccionari, és que són correctes; en cas contrari, es donen a l'usuari perquè les examini.

El component bàsic d'un mecanisme d'aquest tipus és el diccionari. No interessa que sigui molt extens perquè en aquest cas probablement contindrà formes poc usuals i a la vegada allargarà el temps de procés. Si és molt reduit hi ha el perill que la llista de formes no trobades sigui massa llarga. Per a l'anglès es considera adequat un diccionari de 10.000 formes.

Aquest procediment és, en essència, l'utilitzat per la majoria dels programes ortogràfics actualment vigents. Aquests sistemes plantegen, però, problemes diversos: els programes són lents i la llista d'errors no inclou el context en què apareixen.

Els correctors interactius i els diccionaris degudament estructurats permeten una millora en la solució d'aquests problemes.

Els correctors interactius demanen col.laboració cada vegada que no troben una paraula al diccionari. El sistema pot enriquir-se mitjançant diferents modes d'operació per decidir què es fa amb la possible forma incorrecta.

L'estructura del diccionari és també de gran importància sobretot per produir recerques ràpides: per tal de millorar els resultats en el temps del procés s'utilitzen estructures d'arbre, estructures basades en freqüències d'ús, estratègies de dos nivells (formes més freqüents i formes poc freqüents), etc.

La incorporació d'un corrector automàtic és una millora important dels correctors d'errors. El sistema més senzill consisteix en la inclusió d'un diccionari que conté els errors més freqüents amb una indicació de l'ortografia correcta. Un pas endavant en aquest sentit és la generació de possibles formes correctes

mitjançant regles molt simples basades en els errors més habituals: canvi d'una lletra per una altra, omissions, fusió de dues paraules, etc., encara que un procés d'aquest tipus és molt lent i moltes de les paraules que es donen com a alternatives són innecessàries. En aquest cas, un sistema d'anàlisi morfològica reduiria el volum del diccionari. Si es disposés d'un component microsintàctic seria possible el tractament de les concordances sintàctiques i la reducció d'alternatives en els casos d'error.

I.1.3.2.- Sistemes de documentació

Els sistemes de documentació tracten documents textuals, recullen la informació que contenen i la processen. La finalitat pot ser l'alimentació d'una B.D. o una edició més o menys estructurada del document sencer o parcial (1).

Normalment els textos que es tracten en els sistemes de documentació estan altament estructurats i la informació està inclosa en esquemes pre-definits que admeten poca variació. Es caracteritzen per abastar un domini semàntic estret i definit: paraules denotatives, amb poca ambigüitat, etc.

---

(1) H. Rodríguez "Aplicaciones del lenguaje natural" 1987.

Un tipus especial de sistemes de documentació són els **sistemes d'indexació:** la indexació de textos consisteix a guardar la informació textual mitjançant paraules clau per tal de facilitar-ne la consulta i la recuperació. S'utilitza en biblioteques, bases de dades textuals, etc. El procés té dues fases:

a)- selecció de les paraules clau d'un text, els descriptors, i la subseguent incorporació del text a la B.D.:

b)- l'accés posterior a la informació emmagatzemada mitjançant els descriptors.

Normalment el procés de selecció dels descriptors és manual i la seva estructura es redueix a una llista associada al document indexat. La recuperació es fa sobre la base dels mateixos descriptors. Hi ha sistemes que admeten combinacions de descriptors mitjançant operadors lògics.

En els sistemes pre-coordinats, l'estructura del conjunt de descriptors no és una simple llista sinó que s'han explicitat relacions entre ells. Es tracta de les relacions representades en els thesaurus.

Com ja s'ha vist en parlar de les interfícies, en la fase d'accés a la informació és habitual l'ús del llenguatge natural, en canvi la construcció de l'índex

és manual, tant en la selecció dels descriptors com en la seva possible pre-coordinació.

Hi ha intents de selecció de descriptors de manera automàtica basats en factors de tipus estadístic: paraules no funcionals que apareixen amb certa freqüència en el text o bé selecció de les paraules que figuren en un diccionari preestablert. Els resultats no són gaire satisfactoris, excepte en casos de dominis molt restringits.

L'ús de la morfologia i de la microsemàntica, juntament amb una B.D. terminològica, pot donar bons resultats en la selecció de descriptors dels sistemes post-coordinats. Aquests tres components s'articulen de la següent manera: l'analitzador morfològic permet l'estandardització del text, és a dir, l'obtenció del lema de cada forma. Cada lema duu associada la informació semàntica necessària que permet la detecció d'homonimies, polisèmies etc., i la base de dades terminològica dóna les diferents relacions del lema amb d'altres lemes.

La pre-coordinació exigeix sistemes molt complexos de tractament del llenguatge natural. S'han iniciat investigacions en aquest sentit, però no s'ha arribat a cap resultat efectiu.

II.- LA MORFOLOGIA COMPUTACIONAL

La Lingüística Computacional constitueix actualment un nou paradigma que, encara que presenti semblances amb d'altres paradigmes, té els seus propis models.

Normalment els dissenys de Lingüística Computacional es caracteritzen per la seva estratificació i modularitat. Molts dels processos computacionals estan basats en la idea que un procés complex es pot descompondre en una sèrie de processos més simples cadascun dels quals opera amb una certa autonomia.

Les estructures lingüístiques s'acostumen a representar utilitzant una gran varietat de formalismes corresponents als diferents nivells d'estructura: els sons, les paraules i les frases.

L'organització interna de cada nivell és, en certa mesura, independent dels altres nivells. Així, la manera com els sons s'organitzen per formar paraules és totalment diferent de la manera com les paraules s'organitzen per formar frases.

Des d'un punt de vista computacional, el coneixement del llenguatge pot ser analitzat a partir d'un nombre de components separats i de diferents processos que operen amb aquests components. La modularitat del model permet la seva expansió i flexibilitat.

L'esquema de la Fig.1 (1) és d'una seqüencialitat estricta: cada component opera amb els resultats del component anterior i produeix estructures per al següent. És un tipus de representació fràgil, ja que si falla un primer nivell d'anàlisi falla tot el procés, però té l'avantatge que presenta les característiques de modularitat a què he al.ludit.

La morfologia, que és una part dels processos descrits, està estretament lligada, des d'un punt de vista lingüístic, a la fonologia i a la sintaxi. No obstant això, moltes aplicacions de la Linguística Computacional exigeixen que se'n faci un tractament independent. I això per diverses raons:

a)- perquè el reconeixement i categorització de les formes ja és una tasca prou complexa;

b)- perquè el resultat del procés morfològic és imprescindible per al funcionamnet dels altres processos.

---

(1) T. Winograd Language as Cognitive Process. Syntax. 1985.

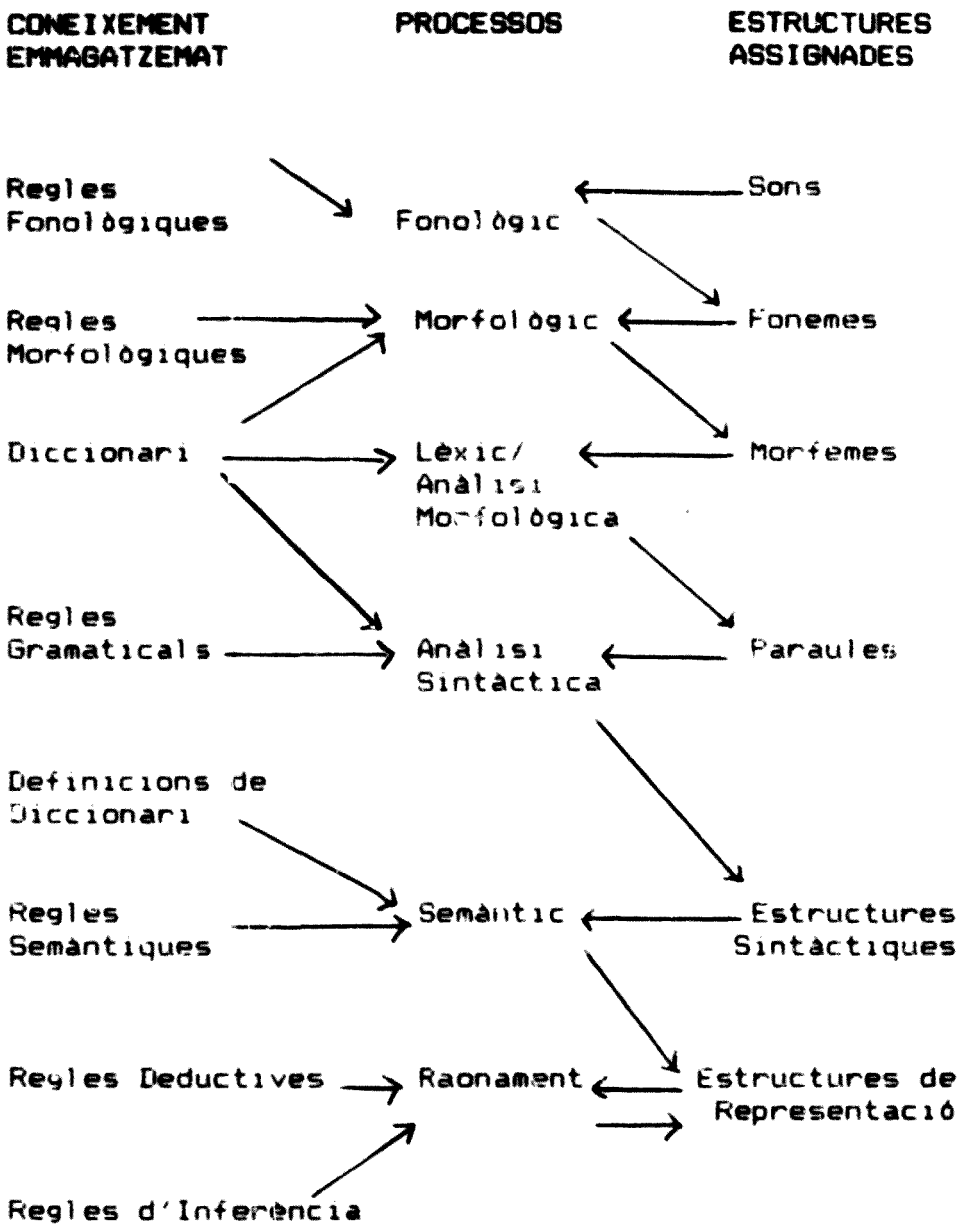

CONEIXEMENT EMMAGATZEMAT
PROCESSOS
ESTRUCTURES ASSIGNADES
Regles Fonològiques
Fonològic
Sons
Regles Morfològiques
Morfològic
Fonemes
Diccionari
Lèxic/ Anàlisi Morfològica
Morfemes
Regles Gramaticals
Anàlisi Sintàctica
Paraules
Definicions de Diccionari
Regles Semàntiques
Semàntic
Estructures Sintàctiques
Regles Deductives
Raonament
Estructures de Representació
Regles d'Inferència

Fig.1

II.1.- Les unitats de l'anàlisi morfològica

S'ha considerat sovint que les unitats d'anàlisi en morfologia són les paraules, i s'entén per paraula la forma lliure amb sentit propi que té una funció dins de l'oració. El concepte de paraula, però, no és tant clar com sembla i això s'observa en diferents casos:

a)- Algunes paraules gràfiques poden contenir més d'una forma:

'del', 'pel', 'cal', etc.

'escalfapanxes', 'tastaolletes', 'llepafils', etc.

b)- Altres vegades la unitat de sentit abraça més d'una paraula gràfica:

- les locucions adverbials : "a cara o creu", "fil per randa", etc;

- les lèxies: "màquina d'escriure", "ull de poll", etc. ;

- les formes perifràstiques dels verbs i els temps compostos.

c)- Els mots d'una llengua no es corresponen exactament un a un amb els mots d'una altra llengua. Expressions que en una llengua es resolen amb un sol mot exigeixen de vegades una frase en una altra llengua per expressar el seu equivalent:

al.: 'Lebensversicherungsgesellschaftsangestellter'

cat.: "empleat d'una companyia d'assegurances",

angl.: "knitwear", "outfitter" "headscarf"

cat.: 'gèneres de punt', 'marxant de confecció' i 'mocador de cap'.

Aquestes qüestions plantegen sovint problemes seriosos a l'hora d'elaborar els mecanismes de tranferència dels diccionaris en els sistemes de traducció automàtica.

El disseny d'un analitzador morfològic o d'un diccionari requereix la definició dels tipus d'unitats ('paraules', 'locucions', etc.) que aquests han d'analitzar i contenir respectivament, i s'ha de fixar molt nítidament el seu abast en funció dels objectius que s'espera aconseguir.

II.1.1.- Classificació dels mots

Des de la perspectiva de l'anàlisi morfològica automatitzada podem considerar que els mots pertanyen a dos tipus de classes: obertes i tancades.

En els analitzadors morfològics, les classes tancades s'emmagatzemen directament en la memòria i se'ls assigna la informació morfològica que els correspon: es tracta de les conjuncions, preposicions, etc.

L'objecte dels analitzadors morfològics són les classes obertes, aquelles que es poden incrementar per préstecs, per creacions o bé per formació de noves paraules a partir de les ja existents.

A la Fig.2 es representen els mecanismes de formació de noves paraules (1):

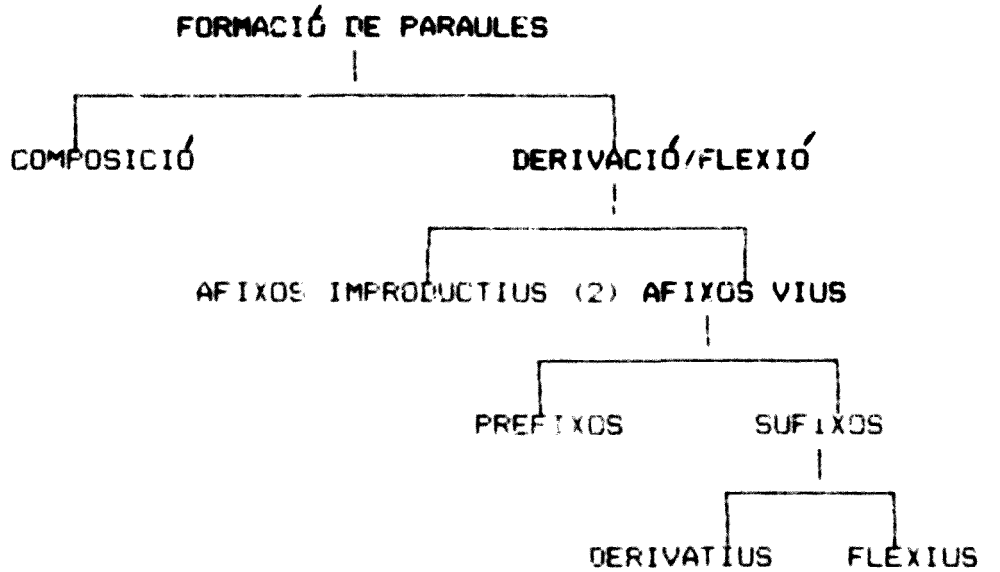


Fig.2

---

(1) N. Cercone "Morphological Analysis and Lexicon Desing for Natural Language Processing", 1978

(2) Cercone utilitza el terme "Dead Affixes", que es pot traduir per 'afixos morts' 'afixos improductius', 'afixos cultes', 'afixos històrics', etc. Creiem que el terme que millor reflecteix el sentit de Cercone és 'improductius', ja que hi ha afixos històrics o cultes que pel fet d'aparèixer com a constituents d'un nombre de paraules considerable s'inclouen com a unitats de l'anàlisi morfològica.

Els diferents sistemes d'anàlisi morfològica ignoren la presència dels afixos no productius, que analitzen integrats a l'arrel. La composició, la derivació amb afixos i la flexió, per la seva productivitat, son l'objectiu dels sistemes d'anàlisi morfològica automatitzada.

II.2.- Objectius de l'anàlisi morfològica

L'anàlisi morfològica és un procés imprescindible per a qualsevol sistema automatitzat que hagi de tractar amb grans quantitats de vocabulari.

Els analitzadors morfològics tenen com a objectiu el reconeixement de les formes d'un text i l'assignació d'informació a aquestes formes, de manera que substitueixen els diccionaris de formes flexives, els quals són d'accés més lent, més costosos de manteniment i que, a més, poden arribar a ocupar molt d'espai.

Els analitzadors morfològics, com veurem més endavant, representen sempre un estalvi, encara que sigui mínim, del nombre de formes emmagatzemades. Els sistemes més simples eviten l'entrada repetitiva de les arrels

comunes a tota una família de paraules o bé l'entrada de les formes flexives d'un mot. Els sistemes més potents resolen l'anàlisi de les formes derivades i dels compostos.

L'anàlisi morfològica ha d'associar informació a les unitats producte del procés. El tipus d'informació varia d'un sistema a l'altre: depèn dels requeriments que se li exigeixen a aquest nivell d'anàlisi. Normalment es tracta de preparar el text per a posteriors tractaments, de tipus sintàctic o semàntic.

Per a les llengües com l'anglès, que es caracteritzen per la seva escassa flexió, l'anàlisi de les formes es resol, normalment, mitjançant el recurs a un diccionari on hi figuren les diferents formes flexives de cada mot amb la informació necessària.

II.3.- Diferents sistemes d'anàlisi morfològica computacional

Presentarem a continuació diversos sistemes d'anàlisi morfològica que, cadascun a la seva manera, donen una solució al problema del reconeixement de les formes flexives i derivades i al de l'atribució d'informació a aquestes formes.

No es tracta d'un estudi exhaustiu de tots els sistemes existents sinó tan sols d'una selecció que pretén mostrar diferents mètodes d'anàlisi, des dels més senzills als més sofisticats. L'objectiu de la seva inclusió és perquè serveixen de contrast amb el sistema d'anàlisi morfològica que hem desenvolupat, amb la finalitat de resoldre l'anàlisi de les formes flexives i derivades amb un mínim d'entrades de diccionari.

Els sistemes que veurem a continuació estan organitzats de menor a major capacitat resolutiva respecte a l'anàlisi de la flexió i en alguns casos també de la derivació.

La resolució de l'anàlisi morfològica pot anar des de la simple construcció d'un diccionari, on són representades totes les formes amb la informació corresponent (és el cas del diccionari de l'LSP que veurem a continuació), a sistemes més o menys

complexos d'anàlisi morfològica. D'aquests, els més senzills consisteixen en la identificació dels components del mot mitjançant la segmentació de caràcters i la confrontació dels resultats amb les dades d'una taula on es troba representat el mot de manera estructurada: els analitzadors de la Universitat de Pisa i de N. Cercone posseeixen aquesta estratègia d'anàlisi. En aquests casos l'estalvi de formes entrades és mínim. D'altres sistemes disposen de diccionaris d'arrels i de sufixos - flexius i derivatius: el procés d'anàlisi consisteix en la segmentació de les formes, la identificació dels segments en els diccionaris i la confrontació dels resultats amb algun sistema de filtratge; aquest pot consistir en la comparació de cada forma amb esquemes de bona formació de mots o en l'aplicació de regles morfològiques que comproven la compatibilitat dels elements resultants de l'anàlisi. La majoria dels sistemes que presentem corresponen a aquesta metodologia: el GRAPHON, el MARS i els analitzadors morfològics de Ben Amadou i Roy Byrd. Finalment hi ha sistemes d'anàlisi morfològica en què l'aplicació de les condicions i la recaptació d'informació es realitza a mesura que es produeix el procés d'anàlisi; per tant, no cal un mecanisme de filtratge posterior perquè cada resultat, des del moment en què ha reeixit, és correcte. L'analitzador de Kimmo Koskeniemmi, així com el que presentem com a tema

d'investigació en aquesta tesi, segueixen aquest tipus d'estratègia.

Per a la presentació de cada un dels sistemes no ens ha semblat adient de procedir a una unificació de la terminologia emprada i per tant hem respectat la terminologia que cadascun d'ells fa servir.

II.3.1.- N. Sager: el diccionari de l'LSP

El lexicó del Linguistic String Project, LSP, (1) és un dels components del sistema d'anàlisi sintàctica de l'anglès desenvolupat a la Universitat de N.York per la doctora N.Sager.

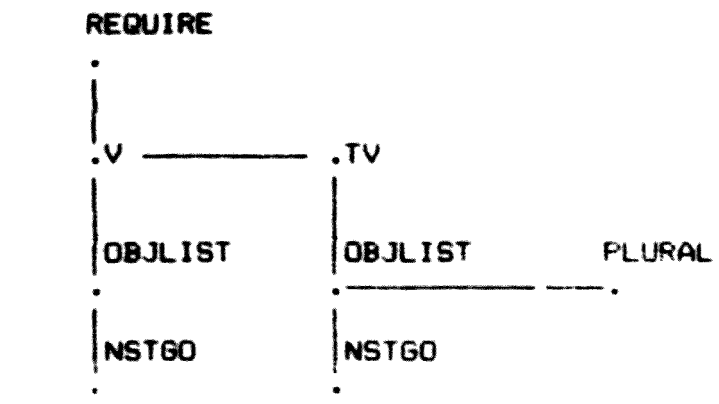
El diccionari conté les definicions de totes les paraules i els signes de puntuació. Les definicions especifiquen les categories de cada paraula i els atributs de cada categoria per a cada forma. Les definicions estan estructurades en forma d'arbre. Cada paraula té com a descendents immediats les diferents categories; i les categories estan caracteritzades per atributs i llistes d'atributs. Vegem el cas de 'require':

```
REQUIRE    V: (OBJLIST: (NSTGO)),   TV: (OBJLIST:
           (NSTGO), PLURAL).
```

---

(1) N. Sager Natural Language Information Processing 1981.

En forma d'arbre:



Hi ha definides 15 categories majors i 150 subcategories. Els atributs poden ser símbols terminals, símbols no terminals de la gramàtica, un símbol d'atribut pre-definit o un 'literal', una forma de la llengua.

Les 'formes canòniques' ajuden a simplificar l'entrada de formes al diccionari: quan una llista de categories apareix molt sovint es pot abreujar mitjançant la definició de formes canòniques. A partir de les entrades:

```
DOG      N:(SING)
CAT      N:(SING)
TABLE    N:(SING)
etc.
```

es pot definir la forma canònica (NSI):

```
(NSI)
N% SINGULAR
```

que es podrà utilitzar en la definició de les paraules en lloc de la llista de categories:

(NSI) CAT

(NSI) DOG

etc.

El sistema disposa, a més, d'altres mecanismes d'ajuda per al manteniment del diccionari; per exemple, les llistes numerades, que permeten simplicar les definicions i relacionar entre si les diferents formes flexives, de manera que aquella part de la definició comuna a totes elles no s'hagi de repetir cada vegada.

Les homografies es poden resoldre mitjançant dues entrades diferents o bé assignant la llista de categories a una sola entrada:

NUMBER N : (SING).
NUMBER ADJ: (COMPARATIVE).

o bé:

NUMBER N:(SING), ADJ: (COMPARATIVE).

La categoria NULL assignada a una paraula indica que aquesta s'ha d'ignorar durant el procés d'anàlisi sintàctica.

Aquest lexicó exigeix l'entrada de les formes una a una i per tant el procés d'anàlisi és inexistent. Les formes estan subcategoritzades amb atributs de tipus sintàctic i semàntic que les restriccions utilitzen

durant el procés d'anàlisi sintàctica.

Thomas E. Ahlswede (1) està treballant en la construcció d'un lexicó relacional per a la generació de textos d'informes clínics mitjançant un sistema expert en diagnòstics en el marc de l'LSP de la Dra. N.Sager. L'estructura de les entrades lèxiques es troba sensiblement enriquida respecte del diccionari que acabem de descriure, però no millora el tractament de les formes flexives. Per a cada entrada lèxica es dóna la següent informació:

1)- L'entrada,

2)- El seu significat representat mitjançant un codi. Les formes homònimes i polisèmiques tenen una sola entrada al diccionari amb les categories i especificacions necessàries.

3)- La classe sintàctica.

4)- El text de la definició entrat per l'usuari. Es tracta d'una informació documental que no es processa.

5)- Una llista d'atributs.

6)- Una definició en forma de càlcul de predicats.

7)- L'estructura de casos (només per als verbs).

---

(1) T.E. Ahlswede "A Tool Kit for Lexicon Building", 1985

8)- Una llista de zero o més relacions amb un o més punters a d'altres entrades amb les quals es connecta a través d'aquesta relació.

Com en el cas anterior, es tracta d'un sistema que no resol l'anàlisi de les formes flexives, ja que figuren totes al diccionari, però que presenta algunes millores importants com és la possibilitat de definir formes bàsiques que es relacionen amb les formes flexives i derivades mitjançant punters, de manera que no cal repetir en cada cas els punts 5-8 de la definició.

II.3.2.- Els diccionaris del sistema de traducció automàtica SYSTRAN

La metodologia lingüística d'aquest sistema de T.A. (1) consisteix en el fet de tipificar els problemes que planteja la llengua font respecte a la seva traducció a la llengua objectiu. El sistema es basa en diversos diccionaris enfocats a la resolució dels problemes d'ambigüitat de les formes.

Cada un dels diccionaris resol l'anàlisi d'un determinat tipus d'unitats. En al sistema anglès -

---

(1) Van Slipe "Description du système de traduction automatique SYSTRAN de la Comission des Communautés Européennes" 1979.

francès cada forma flexiva d'un mot constitueix una entrada diferent. En canvi, el sistema francès-anglès, com que disposa d'un analitzador morfològic, les entrades del diccionari consisteixen només en la part invariable no flexiva dels mots.

A continuació presentem els diversos diccionaris utilitzats per a l' anglès-francès:

Diccionari d'unitermes: els unitermes són seqüències de caràcters separades per espais en blanc. Es tracta de les paraules pròpiament dites, incloses les flexions, abreviatures, signes de puntuació i xifres.

Com que l'anglès presenta un nombre poc elevat de desinències, es realitza una generació automàtica de les formes flexives regulars:

'note'(vb)----------> 'notes', 'noted', 'noting'

'note'(nom)----------> 'notes'

Diccionari d'unitermes d'alta freqüència. Es tracta dels mots més freqüents: signes de puntuació, lletres (de la 'A' a la 'Z'), xifres i els mots més freqüents sense ambigüitat sintàctica o polisèmica ('or', 'an', 'of', 'and', etc.)

Taula de locucions (1). Aquesta taula tracta dos tipus d'unitats: les **substitucions idiomàtiques**(2) i les **locucions**. Les **substitucions idiomàtiques** són expressions compostes que, per raons sintàctiques, es tracten com un pseudo-mot únic, és a dir, com a un uniterme. En concret es tracta de les preposicions i les conjuncions compostes ('on account of', 'as far as'), els verbs compostos ('carry out', 'give rise to'), termes susceptibles d'aparèixer en un o més mots ('light green', 'heat denatured', 'medium term'), i expressions nominals compostes consistents en un verb que modifica un nom ('seal ring').

Aquestes expressions són tractades com a unitermes i poden figurar a l'interior d'altres unitats semàntiques.

Les **locucions** són expressions que no es poden traduir literalment i els mots que les constitueixen solen ser, en la locució, invariables ('on the one hand', 'for many a long year to come').

Aquesta taula permet localitzar les substitucions idiomàtiques unitermes abans del procés de traducció

---

(1) Amb el mot 'locució' traduïm el terme francès 'idiotisme' i el terme anglès 'idiom'.

(2) El terme 'Substitucions idiomàtiques' correspon a l'anglès 'idiom replace' i al terme francès 'substitution idiomatique'.

així com les locucions i donar-los una traducció correcta des d'un bon començament, de manera que s'evitin anàlisis inútils.

Diccionari de les expressions LS. Una expressió LS (Limited Semantics) és una expressió nominal composta de dos o més unitermes, un dels quals, com a mínim, té flexió i no es pot traduir literalment utilitzant el sentit de base de cadascun dels termes constitutius p.e.: 'blast furnance' i 'developing nation'.

Aquest darrer diccionari és accionat per un sistema d'etiquetes a partir del diccionari dels unitermes: cada mot susceptible de constituir una expressió LS remet automàticament a l'entrada corresponent del diccionari de les expressions LS. Això permet un tractament especial de les expressions nominals que necessiten una traducció particular.

Diccionari de les expressions CLS. Les expressions CLS (Conditional Limited Semantics) es caracteritzen perquè es pot atribuir un o més significats a qualsevol dels seus components en funció del context i/o de l'anàlisi gramatical de la frase.

Aquest diccionari, que s'acciona a partir de les regles contextuals o gramaticals, permet tractar les ambigüitats semàntiques sempre que les condicions contingudes en una regla són satisfetes per l'anàlisi

del text font.

La consulta de tots aquests diccionaris es realitza en diverses etapes del procés de traducció. Immediatament després de l'entrada del text es consulten els diccionaris d'unitermes d'alta freqüència, les substitucions idiomàtiques, les locucions i les locucions L.S. Abans de la resolució de les homografies es consulta el diccionari d'unitermes de baixa freqüència i, després de la resolució de les homografies, i per tal de reforçar-la, el diccionari d' expressions L.S. Les expressions C.L.S. es tracten en l'etapa de transferència, és a dir, després de l'anàlisi del text font.

Els unitermes porten associada la següent informació (es tracta sempre, com ja hem dit, de la versió anglès-francès):

a)- La forma de base: el singular dels noms, l'infinitiu per als verbs, etc., en anglès.

b)- Les flexions, generades automàticament si són regulars o bé incloses manualment en el cas de les formes irregulars.

c)- Un codi de prioritat en cas d'ambigüitat homogràfica absoluta, és a dir, que no es pot resoldre amb el programa.

d)- La categoria sintàctica que li correspon.

e)- Les subclasses corresponents a les categories. Així, els noms poden subclassificar-se en comuns, propis, titulació, sigles, etc.; els adjectius en simples , comparatius, superlatius, etc.

f)- Informació gramatical de gènere, nombre, persona i temps.

g)- Identificació dels caràcters en majúscula.

h)- El cas gramatical per als pronoms.

i)- Tipus d'homografia. P.e.: la forma 'light' correspon al tipus [verb conjugat-infinitiu-nom-adjectiu]; 'rose' al tipus [verb- nom- adjectiu], etc.

Per a cada una de les parts del discurs a què pot pertànyer un mot de base o una flexió hi ha una entrada diferent al diccionari amb tota la informació sintàctica i semàntica pròpia de cada ús.

j)- Codis sintàctics que informen sobre les preposicions regides pel mot (en el cas dels verbs, si són transitius o no i el tipus de complementació que regeixen), codis semàntico-sintàctics propis dels noms (concret, abstracte, etc.), dels adjectius i dels adverbis.

k)- Codis semàntics que designen les categories dels mots o de les expressions quan són susceptibles

d'exercir una influència en el significat o en el tractament dels mots veïns. Aquests codis poden utilitzar-se en els programes de traducció o com a condicions de les regles C.L.S. Per exemple, si el mot 'employ' té com a complement directe una paraula amb el codi "professió" ('translator', 'secretary') es tradueix per 'employer', en cas contrari es tradueix per 'utiliser'.

1)- La traducció al francès.

El sistema disposa també d'un tractament especial per resoldre els casos de polisèmia. Al diccionari dels unitermes es dóna preferència a l'equivalent més genèric possible i en segon lloc es tracten les diferents accepcions susceptibles de presentar-se en un corpus com a entrades L.S. o C.L.S.

II.3.3.- C. Subirats: el diccionari electrònic de l'espanyol

A la Universitat Autònoma de Barcelona, sota la direcció de Carlos Subirats, s'està elaborant un "diccionari electrònic de l'espanyol" (1).

El diccionari consisteix en una llista exhaustiva del

---

(1) C.Subirats "Diccionario electrónico del español" U.A.B. 1987

lèxic espanyol (1). Cada entrada té una codificació alfanumèrica que utilitza el paquet de programes de generació automàtica de les formes flexives. Aquests programes utilitzen la informació del codi alfanumèric per generar les formes flexives de la llengua i especifiquen per a cada una de les formes generades les seves propietats morfològiques i la seva relació formal amb les formes de base del diccionari electrònic.

El sistema disposa de dos fitxers de dades. A l'un, hi figuren les desinències verbals classificades pel temps verbal; a l'altre, els algorismes de la conjugació, tant regular com irregular.

El programa de generació de les formes verbals interpreta la codificació alfanumèrica de les entrades verbals del diccionari electrònic per donar la seva conjugació automàtica.

Cada una de les formes verbals generada porta una especificació de la relació que manté amb la forma de base del diccionari electrònic, és a dir, amb l'infinitiu, i a més una especificació de les categories de 'temps', 'mode', 'persona' i 'nombre'.

Actualment s'està treballant en la resolució de la generació de formes nominals i adjectives.

---

(1) C. Subirats utilitza el terme 'espanyol' que hem mantingut en aquest apartat.

II.3.4.- Universitat de Pisa: l'analitzador morfosintàctic de la llengua espanyola (1)

El sistema de lematització automàtica de la llengua espanyola desenvolupat a la Universitat de Pisa per Ratti, Saba, Catarse i d'altres usa, entre d'altres processos, un analitzador morfològic (2).

La lematització, com ja hem dit a l'inici del primer capítol, consisteix en l'agrupament de les diferents formes d'un mateix mot sota una sola entrada, lema, que serveix de punt de referència. Per representar el lema es segueixen les mateixes convencions de la lexicografia, és a dir, els noms són representats per la forma singular; els adjectius per la forma masculina singular i els verbs per l'infinitiu.

Els lematitzadors són útils en diferents àrees de recerca:

- per establir les concordances: trobar tots els contextos en què apareixen les diferents formes d'un lema;

---

(1) En aquest apartat utilitzem el terme 'llengua espanyola' per fidelitat a l'original, a la resta del text utilitzem 'llengua castellana' o 'castellà'.

(2) D. Ratti, A.Saba, M.N. Catarse, G.Capelli "Analizador morfosintáctico de textos en lengua española" s/d.

- per a treballs de tipus estadístic i estilístic: estudis de les freqüències d'una paraula en les seves formes diferents que permeten extreure'n diverses conclusions segons l'objectiu que es persegueix;

- per alimentar diccionaris, ja que si es sotmet un text al procés de lematització es detecten totes les formes noves.

El sistema d'anàlisi morfosintàctica de textos de la Universitat de Pisa és usat en el procés de lematització, primerament, per obtenir l'anàlisi morfològica de les formes i, en segon lloc, per assignar el lema que els correspon.

Aquest procés de lematització es desenrotlla en tres fases:

1)- Fase de pre-processament

2)- Fase de processament morfològic

3)- Fase de processament morfosintàctic

1)- Fase de pre-processament: s'extreuen del text i s'analitzen les paraules sense contingut semàntic i les locucions adverbials (que corresponen aproximadament a un 50% del total de les paraules del text). Aquest procés consisteix a comparar cadascuna de les formes del text amb les formes d'una llista de locucions i de paraules gramaticals. Totes aquelles formes que no es troben en aquesta llista s'han de

sotmetre al procés d'anàlisi morfològica.

2)- **Fase de processament morfològic:** Un cop el text ha estat depurat, es sotmès al processador morfològic. En aquesta fase s'analitza l'estructura interna de la paraula, es reconeixen els seus components flexius i derivatius, s'individualitza el radical correcte i se li assigna la categoria gramatical i el lema corresponents.

L'analitzador morfològic inclou:

a)- una llista de paradigmes nominals;

b)- una llista de paradigmes verbals;

c)- un diccionari de radicals;

d)- una llista de prefixos;

e)- una llista de sufixos;

f)- una llista d'infixos.

a)- **Paradigmes nominals.** Cada paradigma nominal conté d'esquerra a dreta les informacions següents (vegeu la fig. 2):

- un codi numèric que caracteritza el tipus de paradigma associat al radical en el diccionari;

- un codi específic que assenyala cada categoria: A, els substantius; B, els adjectius; i F, els adverbis.

- les desinències corresponents a cadascun dels paradigmes. El gènere i el nombre de les desinències

es determina per la posició que ocupen :

masculi singular = 1a. posició

femení singular = 2a. posició, etc.

L'esquema general dels paradigmes nominals és el següent:

		1	2	3	4
(01)	A	*o	a	os	as
(44)	B	*o	a	os	as
(5_)	F	*mente	-	-	-
(31)	A	*o	-	os	-
(24)	A	-	*-	-	es

Fig.2

b)- Paradigmes verbals. Els verbs s'identifiquen pel codi E seguit d'un número simbòlic que representa el grup de conjugació (1).

La llista de les desinències verbals conté totes les terminacions dels verbs de les tres conjugacions. Les formes de les desinències s'organitzen en sis columnes seguint el criteri de la persona verbal. Una setena columna conté altres informacions sobre les desinències : mode, temps i conjugació. Vegem-ne una mostra: el codi E1 indica que es tracta de les formes

---

(1) Vegeu la Fig. 3

del paradigma flexiu 1 i el número 11 indica que es tracta del present de l'indicatiu:

1	2	3	4	5	6	7
-o	-as	-a	-amos	-ais	-an	E1 11
-ia	-ias	-ia	-iamos	-iais	-ian	E2 12
-é	-iste	-é	-imos	-isteis	-ieron	E4 14

Fig.3

c)- **Diccionari de radicals.** Cada radical conté totes les informacions de tipus morfològic necessàries per a la reconstrucció de la paraula, és a dir, radicals, afixos, categories gramaticals, codis per a l'assignació de les desinències verbals i desinències per al reconeixement de la forma i producció del lema (Fig.4).

Cada lema ha estat segmentat en els seus components: radical, afixos i desinència.

PREF	RADIC	INFIX	SUFIX	PARADIGMA	VALOR PARAD.
de-	ten-			E2	(er)
de-	ten-		-id	B44	(o,a,os,as)
ob-	ten-			E2	(er)
ob-	ten-		-ción	A24	(b, es)

Fig.4

d)- Llista de prefixos. Els prefixos són aquells morfemes que es troben al principi d'una unitat lèxica, verbal o nominal,i que modifiquen el sentit de la paraula primitiva. La llista de prefixos conté "tanto los que tienen valor independiente como las preposiciones 'por', 'para', 'sobre' etc. como (...) aquellos morfemas que no tienen existencia autónoma: 'ob', 'per', 'es', etc." (1).

Aquesta llista inclou també els morfemes que resulten d'un sistema de comparació entre segments que es refereixen a un mateix lexema, com és el cas dels verbs derivats de 'poner': 'o-pon-er', 'su-pon-er', 'pre-su-pon-er', etc.

e)- Llista de sufixos. Són els morfemes no autònoms que s'uneixen al radical i permeten la creació d'unitats lèxiques noves, relacionades gairebé sempre amb la idea expressada pel lexema primitiu.

La llista de sufixos conté, a més dels proposats per la Real Academia, els individualitzats en la fase de segmentació del diccionari considerats com a productius: '-érrim', '-estr-', '-am', '-ern' (p.e.: 'libérrimo', 'terrestre', 'préstamo', 'moderno', etc.)

---

(1) D.Ratti, A.Saba, M.N.Catarsi, G.Capelli, op. cit. pàg. 22

f)- **LLista d'infixos.** Són aquells morfemes que, en la construcció de les paraules, tenen funcions d'enllaç entre els segments productius: entre sufix i sufix (p.e.: 'person-al-i-dad') i entre radical i sufix (p.e.: 'embaj-a-dor' .

Funcionament de l'analitzador morfològic.

Un cop el text ha estat depurat de les paraules sense contingut semàntic i de les locucions adverbials, es guarda indexat alfabèticament en un fitxer.

L'anàlisi morfològica procedeix de la manera següent: llegeix una a una les formes d'aquest fitxer; busca les seves possibles segmentacions: d'esquerra a dreta, per identificar els prefixos; de dreta a esquerra, per identificar els enclítics, les desinències verbals, les desinències nominals, els infixos i els sufixos.

A mesura que l'analitzador va obtenint diferents segmentacions de la paraula, va comprovant segment a segment si aquest correspon a alguna forma del diccionari de radicals. En cas afirmatiu, ha obtingut un radical possible, en cas contrari cal que continuï la segmentació fins a trobar tots els possibles radicals. El fet de no trobar cap segmentació possible que correspongui a un radical del diccionari

significa que la paraula no hi figura i, per tant, que cal incorporar-la-hi.

Es comparen tots els possibles radicals d'un mot amb el diccionari de radicals fins que la longitud de la base és igual a 1. Quan l'analitzador troba el(s) radical(s) corresponent(s) a la paraula que s'analitza, identifica la conjugació o el paradigma flexiu al qual pertany i obté la seva anàlisi morfològica i el lema a què correspon.

La segmentació és bidireccional i la comparació es fa a cada pas de la segmentació. Si l'analitzador trob. una paraul amb dóna totes les possibles anàlisis de la paraul les diferents hipòtesis de lematització.

3)- Fase de processament morfosintàctic: en aquesta fase s'analitzen les formes lèxiques que poden ser objecte de més d'una anàlisi morfològica, es seleccionen les correctes i s'eliminen les falses.

Durant aquest procés l'analitzador es basa en regles de tipus distribucional, fonamentades en el criteri presència / absència de determinades categories gramaticals en el context immediat de la paraula que s'examina.

El resultat que s'obté de l'aplicació de tot aquest procés és un text analitzat sense ambigüitats. El marge d'error i de paraules no analitzades és d'un 20%. Aquests casos residuals es resolen momentàniament de forma manual. Actualment s'està millorant el sistema per reduir aquest percentatge.

L'analitzador de Pisa disposa també d'un sistema de transcripció morfogràfemica consistent en un algorisme que aplica a morfemes particulars determinades transcripcions gràfiques. Aquest mecanisme resol els canvis gràfics dels segments radicals quan es posen en contacte amb algunes desinències i sufixos, com ara les alternances 'pag-/pagu-', 'resolv-/resuelv-', etc.

II.3.5.- Nick Cercone: l'anàlisi morfològica com a part integrant d'un model de memòria computacional

L'estructura interna de les paraules pot jugar un paper important en la sintaxi i en la "producció" de significats.

Nick Cercone, en el seu treball Representing Natural Language in Extended Semantic Networks (1), proposa de representar el significat de les paraules en forma de xarxes semàntiques ampliades, Augmented Transition

---

(1) N. Cercone, 1975.

Networks (ATN) (1).

El lexicó és la via d'accés al significat. Segons Cercone, els mots d'un lexicó han de contenir informació semàntica i funcional. El coneixement funcional de cada mot es resol mitjançant l'anàlisi morfològica, la qual ha de formar part del component lèxic d'un model de memòria computacional, entre d'altres raons perquè:

a)- redueix els costos d'emmagatzematge. És innecessari guardar totes les formes dels ítems lèxics quan existeixen regles ben definides que permeten la segmentació de les formes i especifiquen la formació de paraules normals. L'ús d'una simple rutina pot evitar l'emmagatzemament de gran quantitat de formes d'un mateix ítem;

b)- és un ajut a la interpretació. Un dels resultats de l'anàlisi morfològica és la definició dels afixos que s'adjunten a l'arrel per formar noves paraules. Sovint aquests afixos, en especial els flexius, poden determinar l'ús d'una paraula en una frase concreta o, com a mínim, reduir les seves possibles funcions;

c)- permet formular hipòtesis sobre les paraules noves. Quan en un text apareix una paraula nova , una

---

(1) Vegeu al capítol I, pàg. 22, la nota a peu de pàgina.

anàlisi morfològica preliminar pot ajudar a determinar la funció de la paraula en el context en què es troba a partir de la informació continguda als sufixos. La possibilitat de fer aquestes inferències sobre paraules desconegudes només és possible a partir d'una anàlisi morfològica àmplia i potent.

d)- dóna informació derivada dels afixos: els afixos derivatius modifiquen sovint el significat dels mots. Per exemple, el prefix 'a-' acostuma a indicar "mancança", '-ment' és un adverbi de manera, '-ança' indica substantiu abstracte, '-ador' és agent, etc. El resultat de l'anàlisi morfològica es pot enriquir gràcies a la informació associada a aquests elements.

El sistema d'anàlisi morfològica proposat per N.Cercone no preveu mecanismes de formació de paraules per composició, ni tampoc contempla els afixos d'ús poc freqüent. Tracta tan sols els afixos vius (prefixos i sufixos). Diferencia dos tipus de sufixos, els flexius (no acumulatius i sempre en posició final) i els derivatius (acumulatius, integradors d'una nova paraula, de distribució més lliure). Els sufixos no productius són tractats com a segments units al lexema amb el qual formen una unitat lèxica no descomponible.

El funcionament de l'analitzador de Cercone requereix haver establert prèviament la llista dels sufixos de

flexió, la dels sufixos derivatius i la dels prefixos, i haver fixat una classificació de les categories lèxiques obertes, OCAT, i tancades, CLOSCAT:

OCAT

N.....nom
A.....acció
NM.....modif. nom
AM ..... modif. adjectiu
etc.

CLOSCAT

CONJ .....conjunció
PREP .....preposició
PRO ..... pronom
NEG ..... negació

Cada una d'aquestes categories conté subcategories:

N

NS.....nom singular
NP.....nom plural
COLL...col.lectius, etc.

A

AUX....auxiliar
WILL...futur
TRANS..transitiu
etc.

PRO

NP.....plural
NS.....singular
REL.....relatiu
DEM.....demostratiu
INDEF...indefinit
etc.

L'accés a l'analitzador es realitza mitjançant dues rutines, STEM i CLASS. El procés d'anàlisi es realitza en dues fases. En una primera fase es descomponen les formes. La rutina STEM, que té com a argument la paraula que es vol analitzar, dóna com a resultat una llista amb quatre elements:

(1), la paraula original

(2), els prefixos

(3), l'arrel

(4), els sufixos

Quan un d'aquests elements no s'actualitza apareix NIL, la llista buida.

RUTINA	**Resultat de l'anàlisi**			
	(1)	(2)	(3)	(4)
=(STEM 'RIDDANCE)	= (RIDDANCE	NIL	RID	(ANCE))
=(STEM 'KISSES)	= (KISSES	NIL	KISS	(ES))
=(STEM 'ASLEEP)	= (ASLEEP	A	SLEPP	NIL)

Una segona rutina, CLASS, forneix totes les possibles categories de cada paraula amb les seves subcategories corresponents. Aquesta rutina té com a argument la paraula que es vol classificar i mira si la forma apareix com una entrada lèxica. Si la resposta és negativa contesta: "I don't know the word" (sic); en cas que aparegui entre les categories tancades, dóna el significat que la paraula té a CLOSCAT. Si es tracta d'una categoria oberta, dóna les característiques rellevants que la paraula en qüestió té al lexicó de categories obertes, OCAT. Vegem-ne un exemple:

'drink' (1)

```
(D(R(I(N(K(*(N  ((NIL NS)(S NP)(ABLE NS)(ETTE  DIM)(IE
                  DIM))
                ((0 0) (*DRINK2))
                ((0 0) (*DRINK4)))
                ((ING NS))
                (((0 0) (*DRINK6)))
                ((ER PERS) (EER PERS) (IST PERS))
                (((0 0) (*DRINK5)))
                (SYN DRAFT POTATION BEVERAGE LIQUOR)
                (ID BOOZE HOOCH MOONSHINE))
           (A ((NIL PRES) (S PRES TPS (ING PART))
              (((0 0) (*DRINK1 P1 P2))
               ((0 0) (*DRINK1A P1 P2))
               ((0 0) (*DRINK3 P1 P2)))
              (SYN CONSUME SWALLOW IMBIBE GUZZLE
               TOAST)
              (ID SWIG SOP-UP))
              etc.
```

Si s'analitza la forma 'drinker', la rutina CLASS extreu del lexicó la part rellevant de l'entrada lèxica basant-se en la morfologia del seu argument:

Rutina **CLASS**

```
(CLASS 'DRINKER)
(N (PERS ((00)) (/DRINK5))))
```

i per a 'drinks':

Rutina **CLASS**

```
(CLASS 'DRINKS
(*(N  (NP)( (0 0) (/ (*DRINK2)) ((0 0) (*DRINK4)))
A(PRES TPS ( (0 0) (/*DRINK1 P1 P2)) ((0 0)
(/*DRINK1A P1 P2)) ((0 0) (/*DRINK3 P1 P2))))
```

---

(1) Les expressions *DRINK1 *DRINK2 etc. del llistat fan referència a les diferents interpretacions semàntiques.

Aquest analitzador permet un estalvi en l'entrada de radicals. La flexió i els afixos s'han d'especificar individualment per a cada radical.

II.3.6.- A.Pounder i M.Kommenda: GRAPHON

Un dels problemes més importants que han de resoldre els sistemes de síntesi de la veu a partir de textos de vocabulari no restringit és la derivació automàtica de la pronunciació correcta. GRAPHON (GRAPHeme-PHONeme-conversion) (1) és un sintetitzador de textos de l'alemany que utilitza un analitzador morfològic per obtenir informació sobre l'estructura interna de la paraula, el seu origen, mot genuí o forà, etc. per tal de produir-ne la transcripció fonètica correcta.

L'analitzador morfològic de GRAPHON es basa en un lexicó únic on hi figuren els prefixos, les arrels, els sufixos, etc. Aspectes com la dièresi, l'alternança vocàlica i la deleció de caràcters es resolen mitjançant l'emmagatzemament d'al.lomorfs (p.e.: 'Apfel' / 'äpfel', 'lauf' / 'lief', 'trocken' / 'trockn'). Les formes irregulars com ara les formes verbals 'sein' - 'bin', 'war' - 'wär', etc. tenen entrada pròpia.

---

(1) A. Pounder i M. Kommenda "Morphological Analysis for German Text-to-Speech System", 1986

El diccionari d'aquest analitzador conté aquelles formes més bàsiques que permeten un tractament econòmic de la derivació i de la flexió. Així, al diccionari hi figura la forma 'nam-' en lloc de 'name' (nominatiu singular) perquè permet tractar les formes flexives i derivades.

El criteri que governa la decisió sobre què constitueix una entrada és fonamentalment pragmàtic: s'ha triat sempre la solució que afavoreix el funcionament ideal del sistema.

Les entrades del diccionari són els lemes, és a dir, la representació grafèmica dels morfs, que tenen assignada informació jerarquitzada que permet la seva caracterització fonològica, morfològica i sintàctica.

La informació jerarquitzada conté les dades classificatòries del morf: s'especifica el seu estatus morfològic (arrel lèxica, partícula, morf derivatiu, morf flexiu, etc.), si es tracta d'un terme genuí o forà i les restriccions combinatòries. El lexicó també permet donar informació sobre l'assignació de la categoria sintàctica i, si és necessari, sobre l'esquema accentual.

El procés d'anàlisi es realitza en tres etapes:

1)- Segmentació automàtica de les formes. L'anàlisi automàtica pren en consideració només aquells segments que corresponen a formes del lexicó, de manera que els segments són contigus i no es deixa cap caràcter sense tractar. Mai no es podrà realitzar la segmentació 'mein+un+g' de la forma "Meinung" perquè el segment 'g' no existeix al diccionari.

El nombre d'anàlisis possibles es redueix sensiblement perquè no es busquen noves segmentacions d'una forma quan aquesta correspon a una simple entrada al diccionari. Per economitzar temps de processament s'utilitza una estratègia que dóna preferència als segments més llargs.

2)- Filtratge. En una segona fase s'examinen les segmentacions que s'han realitzat per comprovar la seva conformitat respecte dels principis de l'estructura morfològica de l'alemany segons l'esquema:

[P + S + D + J] # P + S + D + I

on P és una partícula, S és l'arrel, D és el morf derivatiu, I es el morf flexiu i J el morf de lligam.

S'assigna una descripció estructural a la segmentació comparant els trets combinatoris de cada unitat amb la informació que el diccionari assigna a cada morf veí.

Els morfs s'especifiquen d'acord amb unes determinades propietats i amb la selecció de valors per a aquestes propietats que indiquen:

- si es tracta d'un mot genuí o no;
- la seva flexió;
- el seu estatus morfològic, és a dir, si es tracta d'una arrel, d'un prefix, d'un sufix flexiu, etc.
- la seva classe morfològica.

L'especificació d'aquestes propietats és opcional: com més informació es té més restriccions s'aconsegueixen respecte a la fórmula general i també es redueix el nombre de possibles etiquetes.

Si després d'una primera anàlisi no s'obté cap resultat, es repeteix el procés amb una nova segmentació fins que es troben conjunts de trets compatibles.

3)- Finalment s'aïlla l'estructura correcta i se li assigna la categoria sintàctica a la paraula.

En alemany les classes lèxiques estan determinades, en general, per l'últim element del mot (sufixos o elements de flexió), de manera que l'algorisme de classificació utilitza els resultats del procés d'anàlisi per assignar la categoria sintàctica al mot

a partir de l'últim element identificat.

Després del procés d'anàlisi morfològica, cada unitat del text està provista d'una especificació estructural de manera que poden operar sobre ella el component fonològic, sintàctic i prosòdic: per a cada forma del text s'especifica la categoria sintàctica i la descomposició interna. A partir d'aquestes dades el component fonològic determina la pronunciació correcta i genera la transcripció fonètica segons l'A.F.I. (Alfabet Fonètic Internacional).

II.3.7.- M.Meya: MARS

MARS (1) és un sistema d'ajuda a l'usuari que vol obtenir informació d'un Banc de Dades. Aquest sistema permet trobar els termes que poden ser rellevants: a partir de les paraules que l'ususari dóna a l'ordinador i mitjançant un procés d'anàlisi morfològica, el sistema li ofereix totes les paraules que lingüísticament estan emparentades amb l'originària.

El sistema disposa de les dades següents:

---

(1) M. Meya "Análisis morfológico automático del español" 1983.

M. Meya "Análisis morfológico como ayuda a la recuperación de información" 1986

a)- Fitxers invertits de la BD de referència. Aquestes dades s'han extret de diferents corpus: glossari terminològic de telefonia, articles de EL PAIS, textos literaris, etc.

b)- Dades lingüístiques necessàries per realitzar la descomposició morfològica indispensable per a l'anàlisi:

1- una gramàtica d'estats finits;

2- una llista de morfemes del castellà;

3- una llista de les transformacions grafèmiques;

4- una taula de la jerarquia entre les segmentacions alternatives.

c)- Un autòmat d'estats finits.

L'autòmat d'estats finits realitza la descomposició de les formes: agrupa grafies fins que troba al diccionari de morfemes una entrada igual; en aquest cas ha trobat una possible segmentació. Si els trets que el morfema té associats són adequats al tipus d'estat en què es troba l'autòmat, es passa a l'estat següent. La descomposició d'una paraula en els seus morfemes es fa a partir de 15 estats i les transicions entre els estats. Cada estat correspon a una classe de morfemes diferent. Alguns d'aquests estats estan definits com a finals.

El sistema preveu quinze classes morfològiques: pre-prefix, prefix no adverbal, prefix adverbal, prefix denominal, morfema lèxic no verbal, morfema lèxic lligat, morfema lèxic verbal, pre-sufix no adverbal, pre-sufix adverbal, sufix no adverbal, sufix adverbal, flexió no verbal, flexió verbal i enclítics ('adverbal' té el sentit de morfema contigu a una arrel verbal i 'denominal' és el morfema que transforma la categoria nominal de l'arrel).

L'esquema de la gramàtica morfològica (1) és:

prefix /#/ arrel /#/ sufix /#/ flexió /#/ enclític

on /#/ és un exponent de la combinació permissible entre morfemes.

Els sufixos estan subcategoritzats d'acord amb determinades regles de combinació amb els morfemes precedents i d'acord amb la seva possible posició dins de la derivació. Aquesta subcategorització especifica si els sufixos poden anar immediatament després de l'arrel, en posició intermèdia, final o bé si només apareixen sols.

La gramàtica morfològica es basa fonamentalment en dos principis:

---

(1) En morfologia computacional és freqüent l'ús del terme 'gramàtica morfològica' per indicar les regles o els esquemes que expressen les combinacions correctes dels components del mot.

1)- restriccions lingüístiques que deriven de les diferents categories morfosintàctiques que resulten de la combinació de morfemes i,

2)- restriccions sobre la distribució dels morfemes.

Per tal de reduir el nombre de descomposicions possibles d'una paraula es disposa d'una taula on s'expressa la prioritat entre les diferents descomposicions.

En aquest analitzador el procés es realitza en dues etapes: en una primera etapa l'autòmat efectua les segmentacions del mot que es validen segons les regles de la gramàtica, i en dóna les diferents interpretacions possibles. En una segona etapa es realitza el filtratge consistent en la comparació dels resultats amb l'esquema de mot establert per la gramàtica. Finalment se li assigna la categoria.

II.3.8.- Ben Amadou: un sistema d'anàlisi morfològica de l'àrab

Ben Amadou (1) proposa un sistema d'anàlisi morfològica de l'àrab basat en la identificació, per a una determinada paraula, dels seus components: arrel i afixos.

---

(1) B. Amadou " A Comprension technique for arabic Dictionaries: Tha Affix Analysis" 1986

L'anàlisi d'afixos consisteix en la descomposició d'una determinada paraula en els seus elements : prefix, infix i sufix, que són elements redundants, i l'arrel, element no redundant. La descomposició es basa en el criteri derivatiu: moltes paraules s'obtenen adjuntant una combinació d'afixos a una determinada arrel.

Entre les possibles combinacions d'afixos, Amadou diferencia les correctes d'aquelles que no ho són. Les combinacions correctes constitueixen un model morfològic (1). Per a una determinada arrel, el nombre de descomposicions morfològiques possibles depèn de si conté caràcters que poden ser assimilats pels diferents afixos.

La derivació morfològica a partir d'una determinada arrel pot anar acompanyada de transformacions determinades per fenòmens lingüístics com l'assimiliació, la contracció, etc. Les transformacions poden afectar tant l'arrel com els sufixos. Les transformacions morfofonològiques substitueixen un caràcter per un altre sense modificar la longitud de l'arrel; les transformacions purament fonològiques poden eliminar un o més caràcters i per tant la longitud de l'arrel.

---

(1) Utilitzem el terme 'model morfològic' per traduir l'anglès 'morphological pattern'.

L'anàlisi dels afixos es realitza en dos mòduls:

a)- un mòdul de descomposició morfològica,

b)- un mòdul de validacions.

a)- El mòdul de descomposició morfològica permet identificar les diferents combinacions d'afixos; s'efectua en dues etapes: en una primera etapa s'identifiquen els prefixos i els sufixos i, en una segona etapa, els infixos.

b)- El resultat d'aquest primer procés és una llista de descomposicions possibles sobre les quals s'aplica el mecanisme de validació basat en el principi de congruència de sufixos i en el resultat de la consulta al diccionari d'arrels.

La congruència de sufixos es comprova en tres punts:

1)- Compatibilitat entre el prefix (P) i el sufix (S).

2)- Compatibilitat entre el conjunt prefix i sufix (P,S) i l'infix (I).

3)- Compatibilitat entre l'esquema morfològic prefix- infix-sufix (P,S,I) i l'arrel.

Una matriu de congruència d'afixos determina la compatibilitat entre els prefixos i els sufixos: C(P,S).

La compatibilitat entre el conjunt prefix-sufix (P,S) s'obté intersecccionant el codi morfològic (MC) generat per l'analitzador amb el conjunt de codis morfològics associats a la parella (P,S).

Finalment, la compatibilitat de l'esquema morfològic amb l'arrel no és de tipus morfològic sinó semàntic. La detecció d'aquest tipus d'incompatibilitat exigeix donar informació al diccionari sobre els esquemes morfològics no sistemàtics de les arrels (p.ex. sobre les formes derivades).

La consulta al diccionari permet verificar si la paraula analitzada pertany o no al corpus lingüístic. El diccionari juga un paper fonamental en la identificació de l'arrel vàlida, si l'anàlisi, segons un determinat esquema morfològic, genera diferents arrels candidates.

Aquesta anàlisi per afixos ha resultat molt útil en el cas de l'àrab, llengua amb una gran riquesa derivativa. Amb un diccionari de 1.600 arrels es poden analitzar unes 100.000 formes.

El programa de Ben Amadou ha estat incorporat a un sistema de detecció i correcció d'errors ortogràfics.

II.3.9.- Roy J. Byrd: regles de formació de paraules

Roy J. Byrd (1) proposa un pla d'organització del component lèxic per als sistemes de processament del llenguatge natural basat en un diccionari, on es guarda informació idiosincràtica sobre les paraules, i en un mecanisme interpretatiu, que aplica regles morfològiques (II) derivades de les regles de formació de paraules (I).

El sistema de regles morfològiques de Byrd es diferencia dels sistemes basats en la segmentació d'afixos en tres aspectes:

a)- pren en consideració els trets associats a una paraula tant si són una combinació d'informació totalment idiosincràtica com si es tracta d'informació sistemàtica derivada de l'aplicació iterativa de regles morfològiques.

b)- permet observar conjunts complexos de restriccions a l'aplicació de les regles morfològiques. La seva observació fa possible un control molt més estricte de la combinació dels afixos amb les bases.

c)- proposa anàlisis correctes davant la presència de noves formes.

---

(1) R. Byrd "Word Formation in Natural Language Processing Systems", 1983

Aquest sistema permet construir subsistemes lèxics imitant el comportament de les regles de formació de paraules.

I)- **Les regles de formació de paraules** expressen en quines condicions es pot combinar una arrel amb un determinat sufix. Per exemple:

$$+ee \left\{ \begin{array}{l} [\ [+\text{animate object}]\ V\ \ \_\ ] \\ [+\text{transitive}]\ V \end{array} \right\} \begin{array}{l} N \\ [+\text{animate}] \end{array}$$

Aquesta regla es llegeix de la manera següent: el sufix '-ee' es combina només amb verbs transitius amb objecte animat i dóna com a resultat noms animats. Amb aquesta regla es pot derivar 'draftee' de 'draft', però en canvi no es pot derivar *'singee'('sing'no admet objectes animats) ni *'abdicatee' ('abdicate' és intransitiu).

Les regles de formació de paraules estan sotmeses a tres tipus de restriccions:

a)- les restriccions per indicar els límits (1);

b)- les restriccions de subcategorització i ;

c)- les restriccions de selecció.

a)- les restriccions de definició dels límits serveixen per indicar que determinats afixos només poden aparèixer immediatament després d'uns

---

(1) Traducció del terme anglès 'boundaries'.

determirats contextos. Les més profundes estan associades a afixos no neutres que poden alterar la fonologia de la basc a la qual s'apliquen. és el cas de l'exemple que hem vist : '-ee' altera l'esquema accentual. En aquest cas el límit s'assenyala amb el signe '+'. Els límits del següent nivell estan associats a sufixos neutres com ara '#able' que no alteren la fonologia de la base. S'indiquen amb el signe '#'. Els límits més externs estan associats als sufixos flexius i estan marcats amb el signe '-'. Un exemple és el sufix '-ed' per a la formació del participi.

b)- Les restriccions de subcategorització limiten la presència dels afixos a contextos que contenen determinades categories. Per exemple, el sufix "#able" només pot aparèixer amb verbs, no pot combinar-se amb noms (*'elementable') ni amb adjectius (*'temporaryable'). L'analitzador morfològic de Cercone també té en compte restriccions de subcategorització quan comprova que el segment restant després d'haver separat l'afix correspon a una determinada categoria.

c)- Les restriccions selectives especifiquen determinades condicions sobre trets no categorials de la base de l'afix: per exemple, la base de l'afix "+ee" no tan sols ha de ser un verb, sinó que també

ha de tenir els trets [+transitiu] i [+ objecte animat].

Com que el procés de formació de paraules és recursiu, en el procés de formació es poden canviar o afegir nous trets de manera que el mecanisme de selecció es converteix en una eina potent que, a cada pas, en funció dels trets vigents, indica les restriccions pertinents.

La informació del diccionari ha d'incloure, a més dels trets categorials habituals sintàctics i semàntics, informació morfològica, etimològica i fonològica rellevant per als processos de formació de paraules.

II)- **Les regles morfològiques** consten de cinc elements:

a)- el nom de l'afix;

b)- l'especificació dels límits, que assegura l'aplicació de les restriccions corresponents;

c)- un model , que especifica quina és la base per a la regla mitjançant la descripció del sufix que s'ha de treure i el reajustaments posteriors que s'han de fer;

d)- una condició que inclou les restriccions de subcategorització i de selecció sobre la base;

e)- una expressió que proporciona informació sobre

---

(1) Traducció del terme anglès "pattern".

els trets diacrítics i categorials per al resultat de l'aplicació d'una regla.

Els elements d'una regla es combinen per simular el comportament d'una regla de formació de paraules segons la teoria lingüística d' Aronoff (1976) i Selkirk (1982).

Un exemple de regla morfològica seria:

```
+ ion:  ation5*e*   (verb)   (noun +sg +abstr)
b) a)   c)           d)      e)
```

Per analitzar la paraula 'realization' la regla operarà de la següent manera: la posició del marcador de límit indica que es tracta d'una regla de sufix, per tant es verifica el final de la paraula desde la dreta a l'esquema "a-t-i-o-n". La verificació dóna resultat positiu, s'eliminen cinc caràcters i '*' dóna com a resultat la forma 'realiz', que s'haurà de consultar al diccionari. La consulta falla perquè no es tracta d'un mot de la llengua (la consulta hauria tingut èxit si s'hagués tractat d'un mot com ara "relax-ation") i, tal com s'indica a la regla, s'afegeix una '-e' de manera que s'obté 'realize': ara la consulta té un resultat positiu.

Aquestes regles el que manipulen realment són morfografemes més que no pas morfemes, ja que es tracta de la forma escrita de les paraules. Encara que

això és perfectament acceptable per a les aplicacions computacionals, s'ha de clarificar la correspondència entre les regles abstractes de formació de paraules i la seva realització com a regles morfològiques. Així, cal donar una regla per a cada variant gràfica d'un morfema.

En un subcomponent lèxic construït segons el sistema que acabem de descriure, la informació associada a cada paraula és una combinació dels trets inherents que es troben a l'entrada lèxica (la base a partir de la qual es formen els derivats i les formes flexives) i els trets sistemàtics associats a l'estructura de la paraula. Per exemple, si "realize" és un verb transitiu que exigeix subjecte animat, el derivat "realization" serà també transitiu i exigirà un subjecte animat. Aquest sistema permet associar informació a moltes paraules relacionades sense haver-la de donar particularment per a cada una d'elles.

II.3.10.- Kimmo Koskenniemi

K.Koskenniemi proposa un model lingüístic d'anàlisi morfològica implementat computacionalment (1) . El sistema utilitza un algorisme independent de la llengua de manera que el mateix programa pot operar

---

(1) K.Koskenniemi "Two-level model for Morphological Analysis", 1983.

sobre un nombre ampli de llengües, incloses les altament flexives com el finlandès, el sànscrit i el rus.

Aquest model està basat en un lexicó que defineix:

a)- les arrels de les paraules;

b)- els morfemes flexius i esquemes d'alternança no fonològics i;

c)- un conjunt de regles que defineixen fenòmens orientats des de la fonologia. Les regles estan implementades com a autòmats d'estats finits en paral.lel.

La informació que conté el sistema permet procedir tant a l'anàlisi com a la generació de formes.

Aquest sistema pretén ser una alternativa al formalisme de la fonologia generativa. S'inspira tant en aspectes computacionals com en aquells corrents lingüístics que intenten aconseguir models més concrets i psicològicament més reals.

El model de dos nivells proposa regles paral.leles en lloc de les regles successives del model generatiu. No cal definir estats intermedis en la derivació de formes simples: només existeix el nivell lèxic i el superficial.

El lexicó conté només una entrada per a cada paraula

encara que l'arrel estigui sotmesa a diverses alteracions en les diferents formes flexives en què es pot realitzar. Les formes alternants de les arrels es resolen mitjançant dos mecanismes:

a)- el primer consisteix en un sistema de regles associades als morfofonemes de la representació lèxica, que governen la seva realització en la superfície;

b)- el segon mecanisme consisteix en esquemes d'alternança relacionats amb les formes flexives corresponents. Per exemple:

```
hevo    nen/S    "Horse"
```

on el primer ítem, "hevo", és la representació fonològica de l'arrel i l'últim, "horse", és una informació reservada al lexema, en aquest cas la seva traducció a l'anglès. El segon ítem, "nen/S" indica què pot aparèixer a continuació de l'arrel. En aquest cas es tracta d'un esquema d'alternança:

```
nen/S    nen   S0     " ";
         sE    S123   " ";
```

Aquí, els primers ítems són també les representacions fonològiques (la E majúscula és un morfofonema que es realitza com a zero abans del plural 'i'). Els ítems S0 i S123 fan referència a subconjunts de formes flexives.

Gràcies als esquemes d'alternança es poden definir les variants hevonen, hevosen, hevosia, etc. de l'arrel 'hevo'.

Cada una de les regles és com una equació que satisfà o no una determinada forma i una determinada representació fonològica. Per exemple, en finlandès el plural es forma amb una '-i' que canvia a '-j' si es realitza entre vocals. La regla que ho representa és:

```
i
  <===>  V +  -- V
j
```

El signe + és un indicador de límit entre l'arrel i la flexió. La regla determina que una 'i' al nivell lèxic només pot correspondre a una 'j' al nivell superficial si està situada entre vocals.

En el procés d'anàlisi totes les regles funcionen a la vegada com a equacions simultànies. A partir de la representació superficial es busca la representació lèxica com a solució de les equacions.

Les regles estan implementades com un autòmat d'estats finits on les unitats de l'entrada són parells de símbols, un símbol del nivell lèxic i l'altre del nivell superficial. L'autòmat corresponent a la regla que hem vist és:

	V	+	I	I	=	=	(nivell lèxic)
	V	0	I	J	0	=	(nivell superficial)
1:	2	1	2	0	1	1	
2:	2	3	2	0	2	1	
3:	2	1	4	5	3	1	
4.	0	4	0	0	4	1	
5.	2	5	2	0	5	0	

Les línies numerades corresponen als estats de l'autòmat. Les columnes estan encapçalades per parelles de símbols: el signe "V" correspon a 'vocal', "=" correspon a qualsevol caràcter i el zero, "0", correspon a l'absència de caràcters. El signe + indica el límit de l'arrel. L'estat 1 és l'estat inicial i els nombres de la taula denoten les transicions entre els estats. La transició zero indica que es tracta d'una configuració incorrecta. Un exemple de com funciona l'autòmat seria:

Nivell lèxic:

t a l o + i A

Nivell superficial:

t a l o 0 j a

Estat: 1 1 2 1 2 3 5 2

L'altra alternativa, taloia, no pot reeixir perquè la transició a la columna 1 a l'estat 4 és zero.

S'accepta una configuració quan es compleixen totes

les regles. El conjunt de regles, l'autòmat, funciona com un filtre que corrobora la correspondència dels parells formats per la paraula que s'ha d'analitzar i l'entrada del lexicó corresponent. En aquest sentit les formes homògrafes obtenen també totes les seves possibles interpretacions.

A continuació tenim un exemple de l'anàlisi de dos nivells aplicat al finlandès:

```
katolla
katTo$1!A
Roof Subst ADE SG
(='on the roof')
```

La primera línia correspon a la forma que es vol analitzar; la segona és la seqüència d'entrades lèxiques corresponent, d'acord amb les regles; la tercera línia dóna la informació morfològica de les formes i la quarta, la traducció a l'anglès.

II.3.11.- D'altres sistemes: K.Wothke, P.Vergne i P.Pagès

Incloem també en aquesta relació d'analitzadors morfològics alguns sistemes que incorporen mecanismes com els descrits en algun moment del seu procés. El motiu de la seva inclusió no és tant el sistema d'anàlisi emprat com el tipus d'aplicació perquè s'utilitzen.

En primer lloc cal destacar el PRISM , un sistema d'aprenentatge de regles morfològiques sobre la flexió i la derivació, mitjançant procediments de generalització i analogia a partir d'uns exemples que constitueixen el 'corpus d'aprenentatge'(1).

Els exemples consisteixen en parells de paraules entre les quals existeix una relació flexiva o de derivació. La primera paraula de cada parell és la forma font i la segona, la forma objectiu. L'algorisme d'aprenentatge utilitza dos tipus d'instruccions per a la formulació de regles morfològiques:

a)- les instruccions de substitució de prefixos que canvien el començament d'una forma font per tal de generar la corresponent forma objectiu.

b)- les instruccions de substitució de sufixos que canvien el final de la forma font per tal de generar la corresponent forma objectiu.

El sistema funciona de la manera següent: primer es determinen els diferents tipus d'esquemes de substitució d'un corpus d'aprenentatge determinat. A continuació es computa l'índex de freqüència de cada esquema de substitució al corpus.

Si les instruccions que s'infereixen a partir del

---

(1) K. Wothke "Machine Learning of Morphological Rules by Generalization and Analogy" , 1986.

corpus d'aprenentatge permeten generar les formes objectiu correctes a partir de les formes font del corpus, també ho han de poder fer a partir de formes que no pertanyen al corpus. Les instruccions més específiques, les menys freqüents, precedeixen les més generals per tal de garantir que s'aplicaran totes.

El sistema tradueix les instruccions de substitució inferides a partir d'una forma lògica interna a una forma externa.

El sistema no reconeix les excepcions en el corpus d'aprenentage. Si en un corpus sobre la formació del plural aparegués el parell 'goose'-'geese', PRISM inferiria, per a la instrucció de substitució de prefixos, 'goo'--->'gee'/#__ i, per a la instrucció de sufixos, inferiria ' '---> ' '/'ose'__#. Si aquestes instruccions s'apliquen als noms 'good', 'goodness' i 'goon' s'obtindran els plurals incorrectes 'geeds', 'geednesses' i 'geens'. Aquest tipus d'excepcions es poden tractar en una llista a part per tal que no produeixin hipergeneracions incorrectes.

Jacques Vergne i Pascale Pagès (1) treballen en un sistema d'anàlisi del francès que es caracteritza:

---

(1) J.Vergne, P. Pages i I. Paris "Synergy of Syntax and Morphology in Automatic Parsing of french Language with Minimum of Data", 1986

a)- per ser un tractament a diferents nivells en què la sintaxi i la morfologia treballen simultàniament;

b)- per utilitzar un sistema de confrontació (1) de cadenes; i

c)- per l'absència de diccionari.

La informació que s'utilitza durant l'anàlisi està formada per tres tipus de dades:

1)- un petit lexicó d'unes 80 formes que conté els determinants, les preposicions, les conjuncions, i els signes de puntuació;

2)- les regles de deducció morfològica, i

3)- un conjunt d'esquemes de frases nominals.

El procés d'anàlisi consisteix a reconèixer les frases nominals del text i a continuació procedir a l'anàlisi global de la frase. La informació associada a les partícules del lexicó i la que es deriva del processament morfològic serveixen per realitzar aquestes tasques.

El tractament de la morfologia en aquest sistema consisteix a deduir informació a partir de la morfologia de la paraula sense haver de recórrer a

---

(1) Utilitzem el terme 'confrontació' per traduir l'anglès 'pattern-matching'

diccionaris. La llengua sobre la qual s'aplica és el francès. Així, el sufix '- ité' indica qualitat des del punt de vista semàntic i des del punt de vista morfològic es tracta d'un nom femení, singular. Els finals '-isation', '-ification' suggereixen la idea d'acció etc, encara que la morfologia no sempre dóna informació exacta. Per exemple, la forma '-ement' tant pot ser un adverbi (derivat d'un adjectiu) o un nom (derivat d'un verb).

En els quadres de les fig. 5 i 6 hi són representats els analitzadors morfològics que acabem de presentar segons diferents conceptes:

a)- els diccionaris que utilitzen;

b)- el tipus de procés que realitzen i;

c)- el seu ús, per a l'anàlisi o la generació.

També s'hi indica sobre quines llengües s'ha experimentat i el tipus de representació de què parteixen, gràfica o morfofonològica.

		NAOMI SAGER	C. SUBIRATS	RATTI, SABA	N. CERCONE	POUNDER	BEN AMADOU
DICCIONARIS	Diccionari amb formes flexives	X					
	Diccionari codificats		X	X	X		
	Diccionaris amb conjunts de regles					X	X
PROCÉS	Comparació amb taules	X		X	X		
	Regles de partició		X				
	Regles de partició amb llibres					X	X
	Anàlisi morfològica amb autòmates						
US	Anàlisi	X		X	X	X	X
	Anàlisi i generació						
	Generació		X				
	Llengua	ANGLÈS-FRANCÈS	CASTELLÀ	CASTELLÀ	ANGLÈS	ALEMANY	ARAB
	Representació gràfica	X	X	X	X	X	X
	Representació morfofonològica						

		MONTSERRAT MEYA	R. BYRD	K. KOSKENNIEMI	ANTONIA MARTÍ
DICCIONARIS	Diccionari amb formes flexives				
	Diccionari codificats				
	Diccionaris amb conjunts de regles	X	X	X	X
PROCÉS	Comparació amb taules				
	Regles de partició				
	Regles de partició amb llistes	X	X		
	Anàlisi morfològica amb esbrinatge			X	X
	Anàlisi		X		
	Anàlisi i generació	X		X	X
	Generació				
US	Llengua	CASTELLÀ-ANGLÈS ALEMANY	ANGLÈS	FINLANDÈS	CATALÀ-CASTELLÀ
	Representació gràfica	X	X		X
	Representació morfofonològica			X	

II.4.- Les bases de dades lèxiques

Les Bases de Dades Lèxiques (B.D.L.) i els diccionaris constitueixen una àrea d'aplicació de la Lingüística Computacional semblant als analitzadors morfològics, en el sentit que , de manera àmplia, tracten un mateix tipus d'unitats que, en el cas dels A.M., són 'formes' de la llengua i, en el cas dels diccionaris i B.D.L., les 'paraules clau'.

Els analitzadors morfològics, però, estan orientats a la resolució de l'anàlisi de les formes flexives i dels derivats. Aquest procés consisteix en l'assignació de la seva categoria morfològica, l'especificació de les categories menors (gènere, nombre, temps, persona, etc.), la identificació dels afixos, etc. En canvi, les Bases de Dades Lèxiques centren el seu interès en la informació associada a les paraules clau.

Una B.D.L. és un sistema que permet l'emmagatzemament de coneixement, lingüístic i d'altres àrees, de manera estructurada.

La informació continguda a la B.D.L. es troba associada a formes estandarditzades que segons els diferents sistemes s'anomenen 'paraules clau', 'lemes', 'paraules principals' (1), etc.

---

(1) Traducció de la forma anglesa 'headword'. 'Paraula clau' es tradueix 'key word'.

L'accés a la informació es realitza mitjançant aquestes formes estàndard, encara que determinades B.D. permeten, o preveuen, una certa flexibilitat en la manera d'accedir a la informació. En aquests casos és necessari un procés, basat normalment en algun tipus d'anàlisi morfològica, que tradueixi la consulta de l'usuari en una forma comprensible per al sistema.

A continuació presentem breument algunes B.D.L. que poden ser d'interès, bé pel tipus d'informació emmagatzemada, bé per la forma en què hi està estructurada la informació o per l'aplicació en què s'utilitzen.

H.Aiso i M.Isoda (1) proposen un model de Base de Coneixement Lèxic que guarda diferents tipus de coneixement de diccionari en una infrastructura uniforme que proporciona múltiples punts de vista sobre el coneixement emmagatzemat gràcies a un sistema de relacions implícites i d'intèrprets associats.

Aquesta B.D.L. haurà de funcionar com a component d'entorns de coneixement que permetran a professionals de tota mena l'ús de diferents eines de suport per a la preparació i traducció de documents.

L'entitat bàsica del sistema és la 'paraula

---

(1) H.Aiso i M.Isoda "Model for Lexical Knowledge Base", 1986.

principal'. L'usuari ha d'utilitzar una paraula-clau per accedir a la 'paraula principal'. Com que les 'paraules principals' estan estandarditzades i els usuaris no sempre utilitzen la forma estàndard, cal un mecanisme d'estandardització, és a dir, algun tipus d'anàlisi morfològica, que no especifiquen, que permeti accedir a la BD.

Barnett, Lehmann i Zoeppritz (1) proposen un model de diccionari computacional que, juntament amb una analitzador del llenguatge natural, configura el component lingüístic d'un sistema expert en lleis de tràfic.

El diccionari computacional està organitzat com una B.D. relacional per tal d'integrar els diferents aspectes del treball lexicogràfic i permetre un accés ràpid des de l'analitzador.

L'objectiu de la B.D. és donar informació que permeti el processament del llenguatge natural per ordinador. La informació continguda a la B.D. ha de concordar amb la gramàtica i amb els requeriments del processador semàntic.

Al diccionari cada paraula constitueix una entrada;

---

(1) B.Barnett, H.Lehmann i M. Zoeppritz "A Word Database for Natural Language Processing", 1986.

cada entrada té associada informació morfològica i sintàctica. La informació morfològica consisteix en codis de declinació, declinacions alternatives i els sufixos per a cada arrel. La informació sintàctica consisteix en esquemes de règim per als verbs i adjectius segons els treballs de Gross ("Lexicon-Grammar and the Semantic Analysis of French", 1984) i Fillmore ("The Case for Case, 1968). Aquelles paraules que poden tenir associats més d'un conjunt de trets gramaticals tenen més d'una entrada.

És particularment interessant pel tipus de llenguatge de representació utilitzat, el lexicó que proposen Flickinger, Pollard i Wason (1). Aquest lexicó està pensat per suportar la informació lingüística de les Head-driven Phrase Structure Grammars (HPSG) que, d'aquesta manera, poden reduir les regles d'estructura de frase.

Les estructures bàsiques del llenguatge de representació del lexicó són esquemes (2) amb descriptors (3). Els esquemes estan vinculats entre si mitjançant classes i mecanismes d'herència del tipus:

---

(1) D. Flickinger, D. Pollard i T. Wason "Structure Sharing in Lexical Representation" 1985

(2) Utilitzem el terme 'esquema' per traduir l'anglès 'frame'.

(3) Hem traduit l'anglès 'slot' per 'descriptors'.

si F0 és una instància o subclasse d'un esquema més general F1 , la informació guardada a F1 es pot considerar com a part de la informació de F0.

La base de dades lèxiques conté esquemes que especifiquen descriptors de classes de paraules. La classe VERB conté entre les seves subclasses BASE i FINITE. Si s'especifica que l'esquema de la classe VERB té el valor V per al tret MAJOR, aquest valor no s'ha de donar per a cada una de les subclasses perquè hereten aquesta informació. Aquest lligam entre classes és transitiu de manera que la informació es pot heretar a través d'un nombre indefinit d'esquemes intermedis: una instància de la classe FINITE heretarà el tret FINITE per al tret FORM directament de l'esquema de la classe FINITE, i també heretarà el valor V per al tret MAJOR de l'esquema de la classe VERB.

L'ús d'un llenguatge de representació basat en esquemes proporciona una estructura jeràrquica rica per al lexicó, perquè distribueix a través d'aquesta estructura la informació necessària per descriure els ítems lèxics particulars, de manera que cada descriptor d'una determinada classe de paraules només s'ha d'especificar una sola vegada. Així, cal definir esquemes lèxics genèrics per a les categories gramaticals a diferents nivells d'abstracció començant

per l'esquema genèric WORD, dividint-lo i subdividint-lo cada vegada en categories més específiques fins als esquemes que constitueixen les paraules de l'anglès.

L'Istituto di Linguistica di Salerno d'Itàlia, i el Laboratoire Automatique Documentaire et Linguistique (C.N.R.S) de França han treballat durant anys en la construcció de gramàtiques formals de les llengües respectives, l'italià i el francès. La construcció de gramàtiques lèxiques correspon a la primera etapa d'aquest projecte (1).

Una gramàtica lèxica (2) requereix l'estudi d'unes 300 propietats sintàctiques de les paraules. Les dades es guarden en forma de matriu. Aquest tipus de lexicons permeten representar les restriccions de selecció i de subcategorització sense sortir dels límits de la sintaxi formal.

Atès l'alt nivell de desenvolupament de la informació sintàctica dels lexicons de les formes verbals de l'italià i del francès, s'ha pogut realitzar un estudi comparatiu on s'especifiquen les diferents excepcions

---

(1) A. Elia i Y. Mathieu "Computational Comparative Studies on Romance Languages" 1986

(2) Hem traduït el terme anglès 'lexicon-grammar' per 'gramàtica lèxica'.

de cada verb i les diferents estructures regides per cada acepció. Es disposa de dues taules que permeten les comparacions entre ambdues llengües. Una primera taula dóna la correspondència dels lexemes i dels afixos associats a cada lexema. Una segona taula permet la comparació detallada de les propietats sintàctiques distribucionals.

En el marc d'aquesta investigació s'ha dissenyat una interfície d'accés a la B.D., TRANSLEG, que permet escollir el tipus d'investigació que es vol realitzar. Els objectius de TRANSLEG són:

- permetre que l'usuari consulti amb comoditat el banc de dades lingüístiques;

- permetre que el lingüista treballi en una descripció científica d'una o més llengües i que pugui realitzar estudis comparatius gràcies a les informacions lèxiques o gramaticals disponibles.

III.- L'ANÀLISI MORFOLÒGICA: INSTRUMENTS INFORMÀTICS

En aquest capítol descriurem les característiques i el funcionament del sistema informàtic que hem utilitzat per a la construcció de l'analitzador morfològic del català.

Aquest sistema informàtic és en realitat un generador d'analitzadors morfològics (Fig. 1) que permet construir analitzadors de qualsevol llengua, encara que va ser pensat especialment per a llengües que, com el català, el castellà, el francès, etc:

a)- tenen l'estructura del mot analitzable d'esquerra a dreta:

[prefix] arrel [sufix/os] flexió
---------------------------------->

b)- contenen mots l'estructura dels quals permet definir uns components (arrels, afixos i elements de flexió) que presenten un determinat comportament distribucional en el si del mot:

'-s', '-es' , '-os' apareixen en posició final de mot i indiquen plural;

'-ció' apareix al final de mots que tenen la categoria 'nom';

'-em' acostuma a indicar plural i primera persona verbal, etc.

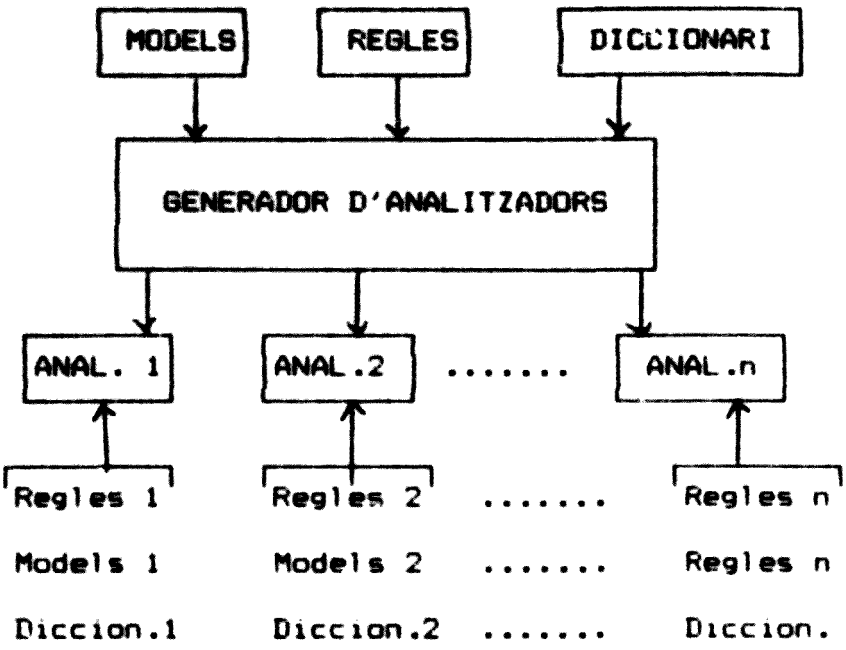


Fig.1

Vam construir aquest analitzador amb la finalitat que, donat un text com a entrada ("input") :

a)- realitzés una segmentació del text en unitats rellevants, de manera que fos possible un tractament posterior sintàctic o semàntic;

b)- associés informació morfològica o d'altre tipus a les unitats resultants de l'anàlisi.

L'analitzador ha estat dissenyat amb la finalitat de permetre l'anàlisi de les formes d'una llengua sense cap mena de limitació pel que fa a l'abast del domini lingüístic.

El problema fonamental que s'havia de resoldre era l'anàlisi d'un corpus molt elevat de formes sense que els costos de temps fossin massa elevats, per tal que l'anàlisi morfològica no alentís la resta dels processos.

III.1.- Elements constitutius de l'analitzador

L'analitzador morfològic es basa en un autòmat markovià, ampliat amb condicions, i que consta dels següents elements (fig.2):

a)- Un **diccionari d'arrels** (1);

b)- Un **diccionari de sufixos** (1) que inclou tant els sufixos flexius com els derivatius;

c)- Un conjunt de **regles** que constitueixen l'autòmat i permeten la concatenació de les arrels amb els sufixos flexius i derivatius.

d)- El conjunt dels **models** en què s'agrupen les arrels i els sufixos segons les seves característiques de flexió i derivació.

e)- Els **atributs morfològics** associats a les unitats diverses dels diccionaris i als models, que serveixen també per expressar les restriccions de les regles de l'autòmat.

---

(1) A l'apartat IV.2.2.1 tractem el tema de la terminologia; de moment utilitzem la terminologia habitual a les gramàtiques.

Les dades lingüístiques estan organitzades internament com una base de dades, amb la qual cosa se'n facilita la consulta i el manteniment.

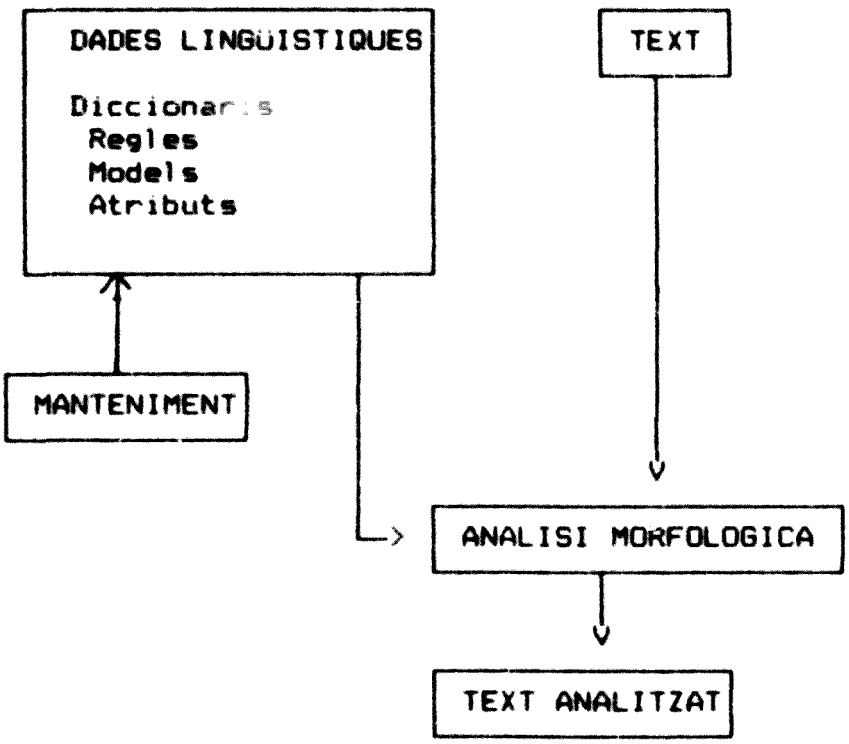


Fig. 2

III.1.1.- Els diccionaris

Internament, l'analitzador diferencia tres tipus d'unitats:

- aquelles unitats que pertanyen a models que tenen l'estat START (1) com a estat inicial i final. Corresponen als anomenats prefixos.

- unitats que pertanyen a models que tenen START com a estat inicial i qualsevol altre estat com a estat final. Corresponen, a grans trets, a les anomenades arrels. Aquestes poden ser 'bàsiques' (RB) 'secundàries'(RS) o 'normals' (RN) (2). La diferenciació entre RB i RS serveix per relacionar les variants ortogràfiques com 'menj-' i 'meng-', 'va-' i 'an-'. Les arrels caracteritzades com a 'normals', RN, no presenten variants ortogràfiques.

L'analitzador tracta els compostos com a noves entrades de diccionari.

- unitats que pertanyen a models que tenen com a estat inicial un estat intermedi i com a estat final un altre estat intermedi o un estat final. Corresponen als sufixos de flexio i derivatius.

---

(1) START és l'estat inicial de l'autòmat.

(2) RB, RS i RN són els codis utilitzats en el diccionari d'arrels per caracteritzar-ne els diferents tipus.

Com es pot apreciar a les fig. 3 i 4, en els diccionaris cada entrada té la informació següent:

1)- Un indicador d'indivisibilitat, INDIV, que es senyala amb el caràcter '*' (1).

2)- El número de la interpretació, No.INT., representat mitjançant un número natural (1, 2, 3,...), indicador de quina interpretació li correspon en el cas que la forma aparegui al diccionari més d'una vegada.

3)- El codi del model al qual pertany, MODEL.

4)- el tipus, TIPUS, d'entrada de què es tracta: com ja hem exposat, si es tracta d'una arrel estarà caracteritzada com a RB, RS o RN; si es tracta d'un sufix estarà caracteritzada com a S;

5)- Una llista d'atributs (2);

---

(1) Sobre aquest indicador, vegeu més endavant III.1.1.1.

(2) Els codis dels atributs que apareixen a les fig.3 i 4 tenen el següent significat:

- l'atribut CAT indica la categoria morfològica i els valors VERB, NOM i ADV, 'nom', 'verb' i 'adverbi' respectivament.

- l'atribut NBRE indica 'nombre' i els valors PL i SG, 'singular' i 'plural' respectivament.

- l'atribut PERS indica 'persona' i els valors 1, 2, 3 si es tracta de la primera, segona o tercera persones.

- TEMP indica 'temps' i PRES, 'present'.

- MODE indica 'mode' i IND, 'indicatiu'.

Diccionari d'arrels:

SEGMENTS	INDIV	No.INT	MODEL	TIPUS	ATRIBUTS
menj-		1	<M1>	RB	CAT=VERB
menq-		1	<M2>	RS	CAT=VERB
flor-		1	<M3>	RN	CAT=NOM
ven-		1	<M4>	RB	CAT=VERB
ven-		2	<M5>	RB	CAT=VERB
de genollons	*	1	<M6>	RN	CAT=ADV

Fig.3

Diccionari de sufixos flexius i derivatius:

SEGMENTS	INDIV.	No.INT.	MODEL	TIPUS	ATRIBUTS
-es		1	<MA>	S	NBRE=PL GEN=MASC
-em		1	<MB>	S	NBRE=PL PERS=1 TEMP=PRES MODE=IND
-es		2	<MC>	S	PERS=2 NBRE=SG TEMP=PRES MODE=IND

etc.

Fig.4

III.1.1.1.- L'indicador d'indivisibilitat

L'analitzador permet tant l'anàlisi de formes lliures, és a dir aquelles seqüències de caràcters situades entre dos espais en blanc, com la de locucions que poden incloure blancs com ara: "fil per randa", "ara i adés" i també la de formes verbals com "hi ha". Per tal d'evitar anàlisis innecessàries per a cada una d'aquestes expressions, com a formes lliures i com a locucions, que donarien el resultat:

fil per randa	CAT=ADV TADV=MANE (Adverbi de manera)
fil	CAT=NOM GEN=MASC NBRE=SG
per	CAT=PREP
randa	CAT=NOM GEN=FEM NBRE=SG

l'indicador d'indivisibilitat evita la descomposició d'una forma en les unitats que la componen: quan en el procés d'anàlisi es troba una forma qualificada com a indivisible no es continuen buscant noves segmentacions possibles. Cal tenir en compte, en l'assignació d'aquest indicador, que les formes de la locució no puguin funcionar com a formes lliures en d'altres contextos: p.ex. "a vessar" no es pot qualificar com a indivisible perquè en la frase "Ha vingut a vessar el contingut del got" no té el valor

de locució sinó que es tracta d'una preposició i d'un infinitiu, en canvi les expressions com ara "de genollons", "sense suc ni bruc" etc. es poden qualificar com a indivisibles sense perill d'anàlisis incorrectes.

III.1.1.2.- El nombre d'interpretacions

Tant en el diccionari d'arrels com en el de sufixos flexius i derivatius, una mateixa forma pot tenir més d'una interpretació. En aquest cas cal donar una entrada al diccionari per cada una de les interpretacions possibles. Per exemple, la forma 'ven-' del diccionari d'arrels s'ha d'entrar dues vegades, una per a l'anàlisi del verb 'venir' i els seus derivats i la segona per a l'anàlisi de les formes de 'vendre' i derivats. Això es possible sempre que les arrels pertanyin a models diferents.

Entre els sufixos flexius i derivatius l'homografia és molt més freqüent. Per exemple, la forma '-es' pot tenir com a mínim quatre interpretacions diferents:

1- Masculí plural de noms ('hom-es');
2- Femení plural de noms (don-es);
3- Segona persona del singular del present d'indicatiu ('meng-es').
4- Plural masculí o femení dels adjectius ('jov-es').

En aquest cas, les limitacions són les mateixes que per al diccionari d'arrels.

III.1.2.- Els models

Les arrels, els sufixos i els elements de flexió s'agrupen en models segons les seves característiques flexives i derivatives.

Els models representen un nivell superior d'organització del material lingüístic que permet representar els comportaments homogenis en la flexió i en la derivació: totes les arrels amb un mateix comportament flexiu estan agrupades en un mateix model, i els sufixos i elements de flexió que es combinen amb un mateix model d'arrels estan igualment agrupats en un únic model.

L'autòmat concatena models, de manera que les combinacions de les arrels amb els sufixos flexius i derivatius no s'han d'especificar arrel per arrel i sufix per sufix, sinó al model al qual pertanyen.

L'analitzador distingeix dos tipus de models, aquells que funcionen en regles que tenen START com a estat inicial, i aquells que tenen com a estat inicial qualsevol altre estat de l'autòmat: són els models dels sufixos i dels elements flexius.

III.1.3.- Els atributs

Els atributs són els portadors de la informació morfològica. Cada atribut pot tenir diversos valors.

Els atributs poden associar-se als models o als elements dels diccionaris, segons es vulgui donar la informació a les entrades dels diccionaris en particular o a totes les entrades d'un model.

Els atributs serveixen també per expressar les condicions de les regles de l'autòmat, segons veurem a continuació (1).

III.1.4.- L'autòmat

L'autòmat consta d'un conjunt d'estats (E), un conjunt de regles (R), un conjunt d'estats finals (F), que és un subconjunt del conjunt dels estats, i un estat inicial axiomàtic, START, que pertany al conjunt dels estats:

$A = (E, R, F, s)$

$F \subset E$

$s \in E$

$E = ($ estats $)$ $R = ($ regles $)$

$F = ($ estats finals$)$

$s = ($ estat inicial $)$

---

(1) Vegeu III.1.4.

Cada regla consta d'un estat inicial de regla (e_i), un estat final de regla (e_f), un model (m) i un conjunt de validacions (W) que és subconjunt del conjunt general de les validacions (V):

$$R = (R_i)$$

$$R_i = < e_i, e_f, m, W >$$

$$e_i \in E \qquad e_f \in E$$

$$m \in M \qquad M = (\text{models})$$

$$W \subset V \qquad V = (\text{validacions})$$

Per tal que una anàlisi es realitzi satisfactòriament, cal que l'autòmat, a partir de l'estat inicial, arribi a un estat qualificat com a final.

L'estat inicial és axiomàtic i s'anomena START, els estats finals els defineix el propi lingüista, així com els diferents estats intermedis.

Si al llarg del procés d'anàlisi d'una forma, l'autòmat no aconsegueix arribar a un estat definit com a final, l'anàlisi no es duu a terme.

Si al llarg del procés d'anàlisi l'autòmat arriba a més d'un estat final, significa que la forma té més d'una interpretació possible i que per tant, el resultat del procés serà les diferents interpretacions que hagi trobat.

El disseny de l'autòmat correspon a un autòmat determinista, és a dir, que un mateix estat només pot funcionar un sol cop com a estat inicial d'un model, encara que pel seu funcionament no és determinista ja que investiga totes les possibles anàlisis que pugui tenir una forma a partir de la informació continguda al diccionari.

Veurem a partir d'un exemple com funciona l'autòmat. Suposem que <ARR1> és un model d'arrels i que <FL1> i <FL2> són dos models d'elements flexius. Les arrels de <ARR1> es combinen amb els elements flexius de <FL1> però no amb els de <FL2>.

Cal donar al model d'arrels un atribut, p.e. FL1=SI, que servirà per expressar la condició de la regla d'elements flexius: d'aquesta manera les arrels de <ARR1> es combinaran amb els elements flexius de <FL1> però no amb els de <FL2>:

model: <ARR1>

atribut: FL1=SI

Sigui START l'estat inicial de l'autòmat; E1, E2, etc. els successius estats intermedis; i F un estat final. Les regles de l'autòmat que realitzen l'anàlisi de les formes correctes són:

R1 START <ARR1> E1

R2 E1 <FL1> F
condició : FL1=SI

R3 E1 <FL2> F
condició : FL2=SI

El pas per la regla R2 només serà possible si en algun pas anterior de l'autòmat s'ha recollit la informació FL1=SI, i el pas per la regla R3 serà possible si s'ha recollit la informació FL2=SI. Per tant les arrels de <ARR1> es combinaran amb els elements flexius de <FL1> i no amb els de <FL2>.

Gràcies a les condicions de les regles s'eviten passos innecessaris de l'autòmat i es disminueixen els costos en temps d'anàlisi així com l'anàlisi de formes incorrectes.

III.2.- El generador

L'analitzador té associat un mecanisme generador que dóna totes les formes flexives i derivades a partir d'una determinada arrel (Fig.5).

El generador permet posar condicions a la generació. Així, per a una determinada arrel es pot demanar que només generi les formes nominals o verbals, o bé les masculines , etc. sempre d'acord amb el codi dels atributs associats.

El generador és una eina molt útil tant per a la construcció de l'analitzador, ja que permet controlar fàcilment el vocabulari que analitza, com per fer estudis sobre el vocabulari, en especial sobre els derivats.

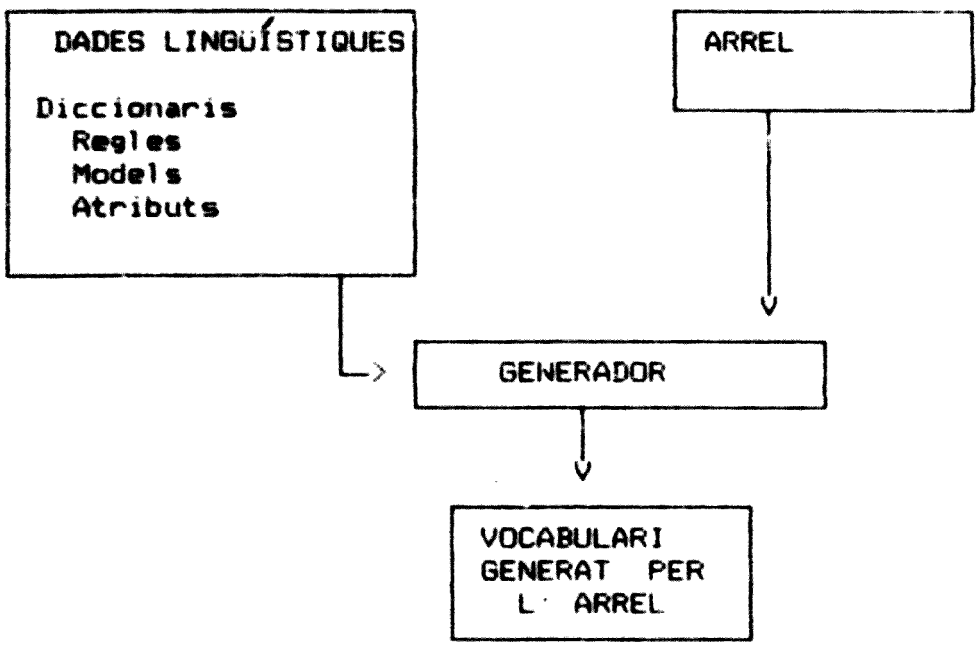


Fig.5

L'analitzador, escrit en FORTRAN, està implementat en un VAX 750.

III.3.- Programes de manteniment

A partir d'una minuta inicial s'accedeix als programes de manteniment dels atributs, de les regles, dels models i dels diccionaris que permeten la seva consulta i modificació.

III.3.1.- Manteniment dels atributs

Els programes de manteniment dels atributs tenen l'objectiu de permetre que el lingüista defineixi les dades corresponents al lèxic. Permeten definir les diferents classes morfològiques , les categories que pot tenir cada classe i els valors que poden prendre cada una de les categories.

Es poden donar de baixa atributs i valors sempre que no intervinguin en alguna regla o estiguin associats a un model o a entrades dels diccionaris.

La consulta dels atributs es pot fer de manera interactiva o bé mitjançant llistats d'impressora.

III.3.2.- Manteniment dels models

Els programes de manteniment de models permeten la definició dels models d'arrels, de sufixos i d'elements de flexió. Els programes permeten l'entrada, la consulta, la baixa i la modificació

dels models.

La modificació de models permet canviar el seu codi, la seva definició, afegir-hi atributs o bé treure'n.

Un programa auxiliar dóna les relacions dels models amb l'autòmat, és a dir les regles en què intervenen, amb l'estat inicial i final de cada regla .

També es poden obtenir per a cada model les relacions del model amb el diccionari: el model amb els seus atributs i les entrades que conté.

III.3.3.- Manteniment de les regles

Els programes de manteniment de les regles permeten que el lingüista construeixi l'autòmat que utilitzarà l'analitzador morfològic. Les funcions que permeten realitzar són:

- la definició de regles,
- l'eliminació de regles,
- la consulta de regles,
- la modificació de regles,
- la generació de la gramàtica i
- la generació del vocabulari definit per la gramàtica.

La modificació de les regles permet augmentar el nombre de validacions o bé reduir-lo així com canviar la seva descripció.

III.3.4.- Manteniment del diccionari

Els programes de manteniment del diccionari permeten donar d'alta , de baixa, consultar i modificar qualsevol de les entrades del diccionari.

El programes de modificació de les entrades dels diccionaris permeten, per a cada entrada:

- canviar el seu model,
- augmentar o disminuir el nombre d'atributs i
- canviar l'indicador d'indivisibilitat.

III.3.5.- Base de dades documental

L'anàlisi morfològica es pot realitzar de manera interactiva o bé en diferit. En aquest darrer cas l'analitzador té connectada una base de dades documental que permet seleccionar l'àmbit dels textos per a l'anàlisi. Un cop definit l'àmbit, s'inicia el procés d'anàlisi i, en finalitzar, el resultat surt per la impressora.