RAMON CERDÀ

LEXICOGRAFIA I INTEL·LIGÈNCIA ARTIFICIAL

0. INTRODUCCIÓ

El contingut d'aquest treball, d'encàrrec, és extraordinàriament extens encara que només comprengui la intersecció de la lexicografia i la intel·ligència artificial. Havent de seleccionar, per raons d'adequació i realisme he triat adreçar-me primordialment a lectors amb formació lingüística. L'exposició està concebuda en bloc com una tipologia de la complexitat de les modalitats lexicogràfiques a partir d'un parell ordenat d'eixos: 1) el caràcter tancat o obert del text i, dintre del segon, 2) l'adopció de criteris sintàctics, semàntics o conceptuals de processament. De passada m'he esforçat a fer veure els vincles i les implicacions de tota mena que hi ha entre la lingüística computacional i la teòrica. No m'he pogut estar d'utilitzar sovint l'enutjós recurs a les notes al peu com un text en paral·lel tot assegurant-me que, en una lògica compensació, la lectura del text pelat en pugui prescindir sense dificultats.

La bibliografia queda reservada a unes recomanacions orientatives al final.

1. L'ESCENARI DE LA INTEL·LIGÈNCIA ARTIFICIAL

1.1 Simular i, si pot ser, emular tot el comportament humà no ha deixat mai de ser la fita prioritària de la informàtica en general. S'hi oposen dues grans limitacions correlatives: 1) la complexitat del comportament i 2) la seva integració. Ara per ara

és impossible reunir en una sola programació global totes les fonts de decisió i de processament que vénen a confegir actes tan senzills en aparença com, per exemple, reconèixer i saludar un amic pel carrer. La simulació s'ha hagut de conformar a actuar sobre unes quantes pautes regulars i a extreure'n models no sols simplificats, sinó també molt més dissociats del que hom voldria. D'aquí ve que la simulació humana global s'hagi fraccionat en sengles disciplines particulars:

(1) (a) la robòtica, que simula el comportament físic,[1]
(b) el processament del llenguatge natural (PLN), que simula el comportament lingüístic, i
(c) la intel·ligència artificial (IA), que simula el comportament intel·lectual.

Seguint un model humà, és lògic considerar que la IA governa les altres disciplines i en crea el marc natural d'integració. D'aquest angle estant podem considerar, doncs, el PLN com una part de la IA i aquesta com una part de la informàtica. També convindrem que si a un sistema se li demanen decisions estrictament lingüístiques –com ara verificar l'ortografia o l'estil d'un text, recuperar cadenes lèxiques o traduir–, el sistema en qüestió pertany a l'àmbit del PLN. Si se li demanen, en canvi, decisions 'pràctiques' –com calcular la direcció i la velocitat d'un mòbil i predir-ne la posició exacta en un temps donat o bé optimitzar la producció i els costos d'una fàbrica d'electrodomèstics–, el sistema pertany a la IA.[2]

Deixant de banda la robòtica i pel que fa a la IA i al PLN, si haguéssim de fer una caracterització molt sumària del decurs històric més recent, valdria a assenyalar unes quantes etapes:

(2) (a) Els informàtics primitius, als anys 50, preveien programes relativament senzills i, com a molt, llargs per construir sistemes capaços de raonar, decidir, sostenir converses i traduir entre llengües. Els fracassos foren espectaculars.
(b) Cap a finals de la dècada del 60, es va comprovar que calia canviar radicalment les concepcions bàsiques i es va demanar l'ajut de psicòlegs i lingüistes per construir programes no trivials d'inferències, sobretot en traducció automàtica (TA). Al mateix temps s'aconseguien avenços substancials i convergents en els camps del hardware, del software i de la representació lingüística.[3] A l'entorn de 1970 es desenvoluparen els primers sistemes experts (3.3) de gran abast enmig d'unes expectatives gairebé utòpiques sobre les possibilitats de la IA.
(c) Vint anys més tard es pot dir que s'han acumulat grans quantitats de coneixement efectiu sobre molts fenòmens del comportament humà tant racional com lingüístic, però manca encara una teoria global psicolingüística i sobretot resultats pràctics d'alt nivell prou atractius per propiciar inversions encara més grans en la recerca bàsica. Es renuncia, si més no momentàniament, a la integració teòrica i es produeix una concentració d'interessos a la bus-

(1) Encara que sol dur incorporat algun tipus de sistema 'intel·ligent' avançat (per exemple, per al reconeixement de formes), la seva característica pròpia és la precisió cinètica i motriu.

(2) En general, tot i que avui dia s'han anat incorporant especialitats originades en disciplines tradicionalment autònomes (lògica, psicologia i lingüística sobretot), la inspiració fonamental i la terminologia ('hardware', 'llenguatge natural', 'implementació', etc.) provenen de la informàtica (vegeu també nota 11). No seria raonable resistir-nos-hi aquí.

(3) Durant aquells anys crucials aparegueren el xip com a substitut del transistor, en el camp del hardware; en el camp del software, els llenguatges de programació –com ara el COBOL, BASIC i FORTRAN, que venien a sistematitzar i, per tant, a ampliar i simplificar enormement la comunicació amb la màquina– i, en el camp de la representació lingüística, la primera versió de la teoria estàndard de Chomsky, que acostava la descripció gramatical als requisits de formalització de la programació informàtica.

ca d'aplicacions pràctiques més realistes.[4] Mentrestant, es forma una immensa xarxa d'activitat en l'àmbit de la lingüística computacional (LC) concentrada sobretot als EUA, el Japó i alguns estats de l'Europa Comunitària. Tot això dintre d'un marc tecnològic extraordinàriament dinàmic on les innovacions poden crear de sobte qualsevol situació sense precedents.

Si s'ha perdut el sentit que tenia deu anys enrera discutir la presumpta correspondència psicològica d'un sistema informàtic no sols és perquè encara no se sap res de substancial sobre els processos psicològics que governen el llenguatge i el raonament en general, sinó també perquè se cerquen resultats més que no pas explicacions; es pretén tot plegat que un sistema tradueixi i en general que una aplicació funcioni, i prou. De fora estant, si hi ha res que distingeixi més clarament la lingüística teòrica (LT) de la computacional és la malfiança d'aquesta última envers les especulacions i l'orientació cap a l'obtenció de productes industrialment apetibles. Aquest caràcter aplicat es fa lògicament substancial en tots els apartats de la LC, i la lexicografia, com veurem, no és cap excepció.

1.2 Com és natural, també hi ha aspectes teòrics dintre la LC, que els tècnics anomenen de 'recerca bàsica' i que no sempre es diferencien nítidament dels que crea la LT. No està gens descartat que cadascuna es beneficiï de qualsevol troballa de l'altra i que fins i tot comparteixin models teòrics sencers. De fet, s'ha especulat bastant sobre la interrelació entre nivells metateòrics, teòrics, formals i procedurals. Així, per exemple, la Gramàtica lèxico-funcional (*LFG, Lexical-Functional Grammar*) de R. Kaplan i J. Bresnan fou concebuda per ser implementada com a mecanisme computacional, però el seu desenvolupament ha estat igualment remarcable com a teoria de la gramàtica. També es pot dir el mateix en alguna mesura d'altres models com ara la Gramàtica d'estructura sintagmàtica generalitzada (*GPSG, Generalized Phrase Structure Grammar*) de G. Gazdar, E. Klein, G. K. Pullum i I. A. Sag, o la seva successora, la Gramàtica d'estructura sintagmàtica orientada al nucli (*HDPSG, Head-Driven Phrase-Structure Grammar*) de C. Pollard i col·laboradors. D'altres models teòrics, en canvi, no tenen rellevància estricta per a la LC, sigui per manca de formalització (per exemple, la glossemàtica de L. Hjelmslev) sigui perquè no és computacionalment accessible (per exemple, la teoria de l'adquisició lingüística de J. Piaget) sigui pel seu abast (per exemple, una descripció interdialectal), etc., etc. Com he dit, però, la LC s'orienta cap a l'obtenció de principis i procediments lingüísticament adequats, explícits i computacionalment efectius. D'aquests principis i procediments se'n diu tècnicament 'formalismes', per oposició a les teories gramaticals de la LT, que cerquen una concepció abstracta o, com a mínim, general del llenguatge. Entre els principals formalismes cal citar, si més no, la Gramàtica d'unificació funcional (*FUG, Functional Unification Grammar*) de M. Kay i la sèrie derivada dels sistemes PATR de S. M. Shieber, L. Karttunen i F. C. N. Pereira.

(4) Pel que fa al PLN aplicat a la lexicografia, hi ha una remarcable concentració d'esforços encaminats a la creació de grans corpus de referència i de diccionaris electrònics de diversos nivells, a la reutilització de recursos lexicogràfics, és a dir a la creació d'estris que puguin recuperar automàticament, o almenys semiautomàticament, la informació continguda en diccionaris convencionals per a diversos sistemes de processament, i a establir les possibilitats de migració de formalismes gramaticals per tal d'integrar i optimitzar models en ús més o menys afins i crear-ne d'estàndards d'àmplia acceptació.

2. LES MANIFESTACIONS BÀSIQUES DE LA LEXICOGRAFIA

2.1 Fins fa relativament poc temps, la lexicografia es limitava gairebé a confegir diccionaris que els parlants nadius feien servir per millorar l'expressió i aprendre el significat d'algun mot desconegut i els estudiants de llengües, utilitzant-ne de bilingües, per traduir. Els diccionaris eren estris auxiliars físicament i temàticament independents de les gramàtiques de les llengües respectives. O, almenys, a les classes de sintaxi o de morfologia ni es miraven, com si fossin camps inconnexos. I tot i que els lexicògrafs s'escarrassaven a reclamar que la feina que feien era veritablement científica i lingüística, el cert és que la seva visió de la llengua era la d'un cos esbaconat amb els òrgans distribuïts pel terra en rigorós ordre alfabètic, cosa que difícilment podia reproduir o suggerir cap activitat vital (compareu a 4.3.2 les figures (19) i (20)).

En la LT dels anys 60 ençà, amb la formalització de la gramàtica, la situació ha canviat dràsticament. El diccionari s'ha convertit en un element cada cop més integrat dins el sistema general, fins al punt que sovint ni tan sols no es pot distingir o aïllar com a mecanisme independent. Prenent denominacions i aparences més o menys diferenciades, el lexicó, el sector que conté la informació lèxico-semàntica de la gramàtica, sol ocupar una posició de partença i de control en tots els models, a vegades amb una preeminència explícita, com en la Gramàtica lèxico-funcional, que he mencionat abans.

En aquest aspecte és força simptomàtica la derivació que N. Chomsky fa de la teoria anomenada de Recció i lligament (*GB, Government and Binding*) on el mecanisme generatiu d'estructures oracionals ja no correspon a regles del component de base, sinó a principis de projecció que provenen de les entrades lèxiques subcategoritzades, és a dir amb estipulacions necessàries per a establir els esquemes sintàctics locals (vegeu 2.3 i sobretot el capítol sobre Lexicografia i Models Lingüístics). Així s'ha optat per una concepció lexicografista de la gramàtica pròxima en certs aspectes a la de molts models computacionals, com veurem successivament.[5]

2.2 Dintre del PLN la prioritat correspon a la informació lexicogràfica. Des del principi, la visió *naïve* característica de les primitives experiències en TA (l'activitat més complexa dintre el PLN) ja era essencialment lexicografista, en el sentit que concebia les llengües en primer lloc com a grans diccionaris complementats per unes quantes instruccions combinatòries, és a dir sintàctiques. Traduir es limitaria, segons això, a establir la correspondència biunívoca entre els mots de dues llengües que componen un text i a arranjar el resultat d'acord amb la sintaxi de la llengua de destinació. Els productes d'aquest procediment mot a mot, però, sovint no deixen ni copsar de què tracta el text així "traduït" per la senzilla raó que la massa lèxica d'una llengua no té estructura discreta (en el sentit matemàtic) ni estable i molt menys guarda una relació isomòrfica amb la massa lèxica d'una altra llengua. Qualsevol diccionari bilingüe de butxaca ho mostra prou bé; més encara, la qualitat dels diccionaris bilin-

(5) Curiosament, en suprimir les regles sintàctiques s'invalida el concepte primigeni de gramàtica generativa introduït precisament pel propi Chomsky. Cal afegir, arran d'aquesta aproximació estratègica, que la teoria estàndard es va mostrar molt poc apta per a la implementació computacional i que el mateix Chomsky sempre ha sostingut no haver tingut mai en compte concepcions o fites computacionals.

gües es basa precisament a magnificar i descriure escrupolosament la incongruència lèxica entre dues llengües qualsevol.

Cal advertir, però, que la gran majoria d'aplicacions més perfeccionades no han canviat pròpiament el disseny lexicografista, sinó que n'han identificat les principals mancances i han incorporat un criteri composicional, oracional o textual. En tot cas, la diferència és força transcendent, com veurem més avall, ja que aconseguir que una expressió sencera en la llengua A correspongui d'una manera plausible o útil a una altra expressió sencera de la llengua B requereix un procés de formalització analítica i sintètica d'una complexitat molt variable.

I trobar el camí més curt o, si es vol, el procediment més barat i efectiu representa tot el que es proposa en definitiva el PLN aplicat a la TA i en general tota aplicació lingüística de la IA.

3. TIPOLOGIA DE LA COMPLEXITAT

3.1 El concepte d'optimització o de solució minimista –equivalent a la 'navalla d'Ockham' clàssica– és crucial en informàtica. Aplicat al PLN, representa l'afany de salvar amb el mínim esforç possible la complexitat provinent de tres paràmetres bàsics i teòricament independents: 1) l'abast temàtic del text, 2) la indeterminació del flux de la informació i 3), quan hi ha algun procés de transferència interlingüística, la distància tipològica que separa les llengües implicades (vegeu 4.2 (10)). Aplicat a la IA, representa enfrontar-se a la complexitat 1) de la interacció entre l'usuari i el sistema i 2) de l'execució dels requeriments implicats.

Com dèiem abans, els models en PLN són lexicografistes almenys en dos sentits: 1) perquè sempre es parteix i/o s'acaba en un lexicó, i 2) perquè, sigui quina sigui la finalitat del producte, el procés s'inicia en un text escrit o oral i la primera operació 'intel·ligent' que s'executa sobre ell és de naturalesa lexicogràfica (tot i que el primer escorcoll pot ser purament gràfic o fonètic). Podem convenir que el grau de complexitat d'un sistema computacional determinat pels paràmetres del paràgraf anterior promou la necessitat (no necessàriament correlativa amb cadascun d'ells) de recórrer a mecanismes addicionals més o menys distingibles, a part dels lexicogràfics: el més immediat és essencialment sintàctic, el següent, semàntic, un tercer lògic i finalment un de 'coneixement del món'. Ara bé, tot i que la incorporació creixent d'aquests mecanismes augmenta el grau de complexitat i de poder executiu dels sistemes, el principi de l'optimització fa que es tendeixi a utilitzar sistemes adequats a cada aplicació, vetllant que el preu de l'esforç no ultrapassi el valor del resultat. No té sentit, per exemple, incorporar un mòdul sintàctic a una aplicació que hagi de distingir només entre una dotzena d'opcions significatives.

Cal afegir, d'altra banda, que en la pràctica els models que exemplifico a continuació com a il·lustració dels graus i modalitats de complexitat computacional, en especial els més elaborats, sovint no es donen en forma pura o absoluta en la realitat, sinó en alguna combinació entre ells.

3.2 La producció computacional d'estris lingüístics més simple de totes és purament lèxica i es troba en la presentació d'alternatives binàries de tipus *Sortir? (S/N)* o bé *Vol resguard? (S/N)*, força típiques en els 'menús' de programes d'ordinador i de caixers automàtics, respectivament, on s'insta l'usuari a pitjar la tecla adient representativa de "sí" o "no". L'abast temàtic del text es redueix aquí a tres mots, comptant la manca de resposta. El sistema es limita a comprovar, a través d'algun codi intern de tipus ASCII, amb quina de les tres possibilitats previstes al seu lexicó coincideix la tria de l'usuari. En general, aquesta comprovació lèxica sense anàlisis prèvies ni ulteriors s'anomena 'correspondència de patrons' (*pattern matching*) i s'aplica en textos tancats, és a dir sempre que hi ha un abast temàtic limitat pel propi sistema, en el sentit que la comunicació entre l'usuari i el sistema segueix un flux preestablert –però no necessàriament reduït, ja que pot comprendre un conjunt de centenars de menús i submenús jerarquitzats (com, per exemple, en el WordPerfect)–. No cal que el patró sigui un mot realment existent en una llengua natural. En els exemples de més amunt el patró es redueix a [ϕ['s'/'n']], com hem vist, i en d'altres pot equivaler a una seqüència complexa com ara el nom i cognoms de l'usuari o l'ordre d'imprimir un tros seleccionat d'un text. Les aplicacions de la correspondència de patrons són força populars i diversificades, ja que, a part dels menús informàtics a què hem al·ludit, es troben en productes com ara els traductors de butxaca, capaços d'establir correlacions entre milers de mots o frases fetes d'un bon grapat de llengües (aquest límit és purament comercial) amb o sense síntesi de veu incorporada per simular la pronúncia figurada del resultat, en els reconeixedors de veu per telèfons sense mans o per consultes bancàries o en els 'correctors ortogràfics' de molts processadors de text. I encara, dintre de sistemes molt perfeccionats de TA (vegeu 4.3), s'utilitzen també procediments anàlegs de correspondència lèxica entre termes tècnics gràcies a la seva equivalència interlingüística estricta.

3.3 Una família més complexa de productes computacionals en PLN essencialment lexicogràfics són les interfícies home-màquina en llenguatge natural per a moltes bases de dades i alguns sistemes experts senzills. Tant la base de dades com el sistema expert consten d'un fitxer amb informació ordenada i un mecanisme incorporat per establir relacions específiques d'informació. Com veurem després, la diferència fonamental està en la simplicitat, la integració i l'orientació de l'algorisme relacional del mecanisme. En una base de dades convencional es pretén recuperar informació factual entre les dades ja contingudes en la base; aquesta informació és bàsicament lingüística i

sovint consisteix en la localització i recuperació de cadenes específiques de caràcters, tècnicament mots. Entre els exemples més típics hi ha la guia telefònica, el catàleg d'una biblioteca o el Diari Oficial de la Comunitat Europea. En un sistema expert es pretén, al seu torn, obtenir o, millor, simular informació nova conjuminant les dades ja contingudes en el sistema amb les que li forneix l'usuari. Es tracta ara d'informació conceptual, que requereix la creació d'un cert món possible amb objectes coherents (vegeu 5.2). Els sistemes experts es construeixen a partir de dades, anomenades bases de coneixement, jeràrquicament emmagatzemades per especialistes, o experts, en una matèria (medicina, geologia, enginyeria, finances...) juntament amb un algorisme relacional, anomenat motor d'inferència, capaç d'establir un raonament dialèctic amb l'usuari i finalment de formar un criteri, posem per cas, sobre un diagnòstic mèdic, la probabilitat de trobar petroli en un determinat sòl, les expectatives d'èxit d'una certa inversió o les continuacions imprevisibles d'un conflicte armat a gran escala.

La principal utilitat d'aquelles interfícies, genèricament anomenades 'gramàtiques semàntiques' rau en evitar que s'hagin d'aprendre o memoritzar codis o llenguatges només intel·ligibles per la màquina i en canvi permetin a l'usuari recuperar o obtenir informació a través d'un diàleg interactiu, és a dir de consultes bàsicament no limitades en la seva formulació ni ordenació. La gramàtica semàntica comporta un cert tipus d'anàlisi essencialment orientat cap a la descoberta de 'mots claus' enmig d'estructures més o menys diverses i complexes. El principi es basa en el supòsit que l'usuari pot triar entre una munió d'estils lingüístics segons les seves habilitats o preferències, però que, en qualsevol cas, farà servir un grapat força reduït de mots precisos per descriure l'objecte central de la seva consulta. L'escassa variabilitat d'aquests mots permet confegir un sistema que els localitzi i arbitri una interpretació de la petició global només a partir de la seva co-aparició.

Imaginem que un lector d'una biblioteca general vol saber què hi ha publicat sobre els índexs d'escolarització a Espanya durant els últims deu anys. Com serà una aplicació teòricament mínima d'una gramàtica semàntica? Un sistema així es compon bàsicament de quatre mòduls:

(3)
- (a) un conjunt estrictament jerarquitzat de camps de contingut (inspirats, per exemple, en la classificació bibliogràfica decimal) representats per sengles mots claus
- (b) un conjunt de possibles alternatives lèxiques i morfològiques a cada mot clau de (a)
- (c) un conjunt de regles de detecció de mots claus i de relació entre ells
- (d) el conjunt de publicacions catalogades amb indicació acurada del contingut mitjançant els camps de la llista (a)

La tercera llista és, de fet, una gramàtica i consisteix en una col·lecció de regles que executen tres funcions específiques: 1) identifiquen els mots claus a través de les seves

possibles alternances expressives previstes en (b) i eliminen la resta de mots, 2) estableixen les possibles relacions de sentit entre els mots claus i els camps de contingut adients i 3) componen una llista de les publicacions amb la reunió o la intersecció dels camps de contingut seleccionats.[6]

Vejam, doncs, com funciona. En primer lloc, el sistema negligirà tot allò que no faci referència als mots claus o a les seves alternatives. Coses com, per exemple:

(4) (a) Quines obres hi ha sobre ...?
(b) Què hi ha de ...?
(c) Voldria saber si s'ha editat res sobre ...
etc.

Els mots claus de la consulta són, en canvi, 'escolarització', 'Espanya' i 'període d'anys' en les seves possibles modalitats expressives. Respectivament:

(5) (b) escolaritat, escolar, escolaritzat, estudiant, estudiantil, que estudia, que rep educació, discent, alumne ...
(c) Estat, estat espanyol, espanyol, estatal ...
(d) deu anys, decenni, dècada, període ...

tot tenint en compte les possibles combinacions i alteracions morfològiques que en definitiva poden donar lloc a construccions tan diverses com les següents:

(6) (a) Desitjaria obtenir dades sobre la [població estudiantil] [espanyola] [de 1983 ençà]
(b) Quines publicacions teniu sobre l'[educació] a l'[Estat] dels [últims deu anys]?
(c) Durant el [present decenni], què hi ha sobre la [gent que estudia] a [Espanya]?
etc.

La intersecció dels camps 'escolarització', 'Espanya' i '1983-1993' donarà presumiblement una sola interpretació força congruent amb el que cercava l'usuari. Noteu de pas que, tal com està dissenyada aquesta gramàtica semàntica, en teoria almenys també s'obtindrien resultats anàlegs als de (5) si s'escrivia la consulta a base de:

(7) (a) L'espanyola, quan estudia, estudia de debò, de 1983 ençà
(b) Espanyols i estudiants, molts i grans l'últim decenni
(c) Decenni últim el grans i molts estudiants i espanyols
etc.

és a dir sempre que s'hi incloguessin rèpliques adients dels mots claus.

(6) En el nostre exemple, assignaran la referència als últims deu anys al tema (a l'escolarització espanyola) i no pas a les publicacions mateixes i establiran si la relació entre determinats camps és d'unió o d'intersecció.

Podem concloure que, al marge dels mecanismes morfològics i combinatoris (i per tant sintàctics) auxiliars, el sistema funciona amb principis essencialment lexicogràfics semblants als dels diccionaris electrònics de baix nivell, on els valors de les entrades són fixos. Seguint un símil convencional, els mots claus fan de morfemes biunívocament relacionats amb els camps de contingut, equivalents a semes. I el diccionari més extens (la llista (3) (d) de més amunt), el catàleg de publicacions, consta d'entrades lèxiques subcategoritzades per un conjunt ordenat de semes o camps de contingut.[7]

3.4 El següent grau de complexitat es presenta en el text obert, quan l'abast temàtic de l'aplicació queda indeterminat. No es tracta pròpiament d'una mera ampliació lineal, com si s'afegissin volums a una obra, tot i que això també afavoreix que les gramàtiques semàntiques siguin insuficients. La indeterminació temàtica es dóna quan l'usuari pot triar entre una sèrie oberta –o prou ampla– d'expressions i els valors lèxics dels 'mots', que en els models anteriors eren fixos, passen a dependre del context. En no poder negligir cap segment del text, cal analitzar-ne l'estructura sencera i produir així una nova dimensió composicional, sintàctica i/o semàntica, si més no. Això comporta afegir al lexicó una gramàtica i diversos mecanismes que creïn alguna representació canònica del text inicial capaç de ser utilitzada adequadament per intercanviar informació amb usuaris humans o amb altres sistemes o convertida en text d'una altra llengua. En el primer cas, podria tractar-se de bases de dades i sistemes experts amb moltes variables, correctors ortogràfics avançats i correctors d'estil –que requereixen analitzar, com a poc sintàcticament, l'estructura de textos no predeterminats– i sobretot en la traducció automàtica.[8] En condicions normals això faria, per exemple, que les distintes versions que hem assenyalat a (6) i (7) fossin considerades no equivalents i en particular que (7) (c) quedés rebutjada per agramatical.

Un cop accedim a aquest tipus màxim de complexitat en PLN ja es planteja obertament la simulació sencera del parlant humà en totes o almenys en algunes de les seves funcions lingüístiques. Tothom accepta que hi ha diferències fonamentals –possiblement insalvables– entre el processament lingüístic humà i el computacional. El parlant humà utilitza en paral·lel, és a dir alhora, fonts d'informació tant lingüístiques (coneixement d'estructures fonològiques, morfològiques, sintàctiques, lèxiques, etc.) com extra-lingüístiques (coneixement propi i compartit del món ambiental, del tema que es tracta, la seva coherència i implicacions, etc.). Noti's que aquesta integració informacional representa un mitjà potentíssim per recuperar moltes de les mancances i deficiències que es produeixen en el curs de tota comunicació. Al seu torn, la gran majoria de sistemes computacionals d'avui dia 1) només poden comptar amb informació estrictament lingüística, i 2) incorporen aquesta informació en forma seqüencial o modular, és a dir per un costat fonològica, per un altre sintàctica, per un altre lèxica... seguint algun ordre preestablert (vegeu 4.3.2 (21)). A més a més, el parlant humà

(7) Avui en dia s'han desenvolupat formalismes dintre d'aquesta classe, intermèdia entre la correspondència de patrons i els *parsers* (vegeu nota 11), del tipus anomenat 'SQL' (*Structured Query Language*) que vénen a optimitzar el diàleg interactiu entre usuaris i màquines a base de llenguatges reduïts per selecció de necessitats i que són capaços de processar quantificacions i variables contextuals força més complexes de les que he exemplificat abans (per exemple, ordres que demanen classes d'objectes diferents en alguna característica a les que s'han obtingut en una consulta anterior).

(8) Noteu que, sense control sintàctic, un corrector ortogràfic no detectarà cap raresa en una seqüència com ara *el flors blaus*.

combina les distincions discretes (o digitals) de la fonologia i la morfologia amb distincions contínues (o analògiques) en multitud d'aspectes primordials com ara la modulació de la veu i la semàntica.[9] En canvi, tot i la insistència en l'adaptació de tècniques de raonament aproximat en sistemes experts, els resultats encara no s'acosten significativament als humans.

Aquestes diferències –que desemboquen en la manipulació d'informació redundant, per part dels humans, enfront d'una informació fortament deficitària, per part dels sistemes computacionals– han suscitat la necessitat de cercar tècniques imitatives o bé alternatives. La possibilitat d'incorporar coneixements del món a un sistema depèn molt dels nostres coneguts factors de complexitat (sobretot l'abast temàtic i la indeterminació), però per a molts especialistes és una tasca purament utòpica, ara per ara, al marge d'àmbits d'aplicació molt restringits. Hi ha qui dubta fins i tot de la possibilitat de construir diccionaris electrònics coherents i efectius, tant multilingües com monolingües, que ultrapassin el vocabulari de terminologies tècniques i algun grupet de mots d'una sola accepció (si és que n'hi ha cap). Tanmateix gairebé tots els sistemes més vàlids en IA i PLN es fonamenten en el supòsit que tot això no sols és possible, almenys en teoria, sinó que els formalismes del futur seran essencialment diccionaris integrats (amb informació morfològica, sintàctica, semàntica, etc.) universals.

A continuació ens concentrarem en les modalitats i tècniques lexicogràfiques que s'utilitzen en aquest últim grup més complex i avançat de sistemes computacionals.

(9) L'extraordinària potència semiòtica del parlant li permet reconstruir o, si més no, establir hipòtesis sobre el sentit i la significància empírica de textos (orals, escrits...) encara que siguin molt fragmentaris, incorrectes o anòmals, des d'un simple gargot fins a un missatge críptic, passant per les més alambinades figures de dicció, entre elles, per exemple, la ironia. Sense esforç aparent, el parlant pot canviar radicalment de registre i de tema (o totes dues coses alhora), utilitzar més d'un codi en paral·lel i construir interpretacions diverses sobre un mateix text. El parlant humà, a més a més, no fa un aprenentatge procedural de la llengua materna (assimilant prèviament una col·lecció de regles), sinó declaratiu, en el sentit que infereix, a partir d'un nombre limitat de fets de parla, les estratègies generals que governen la creació de nous fets de parla que mai no ha sentit abans. Com veurem, tot això no fa sinó il·lustrar la indeterminació o, per ser més exactes, la virtualitat essencial dels significats lèxics en l'ús real de les llengües naturals.

4. LEXICOGRAFIA I FORMALISMES EN PLN

4.1 Com dic, la necessitat de recórrer a una anàlisi combinatòria de les expressions assoleix un nivell de complexitat diferent. Des de l'angle de la lexicografia, es caracteritza perquè desapareix el supòsit que els mots tenen un valor fix i en prenen un de determinat pel context, tal com s'esdevé precisament en la parla real. És el fenomen genèricament conegut per 'ambigüitat', que cobreix un seguit ben diferenciat de manifestacions i graus possibles:

(8)
(a) **ambigüitat lèxica:** *Els lingüistes computacionals són decididament* ***rars*** (pot ser en el sentit d'"escassos" o d'"estranys")
(b) **ambigüitat sintàctica:** *La va veure* ***tot passejant*** (no se sap qui passejava, si el subjecte o l'objecte de *veure*)
(c) **ambigüitat fonològica:** *Allà hi ha* ***les ones/les zones*** *més perilloses* (en ambdós casos amb una mateixa pronúncia)
(d) **ambigüitat referencial:** *Va treure la mosca de la truita i se* ***la*** *va menjar* (no se sap què va menjar exactament, si la mosca, la truita o alguna altra cosa referida prèviament)
(e) **ambigüitat pragmàtica:** *plat d'arròs, plat de test, plat de postres* (només

el coneixement del món ens fa interpretar que l'arròs és el contingut, el test la matèria i les postres la funció del plat, però res d'això no es dedueix de la sola estructura gramatical)
etc., etc.

Cal comprendre que l'ambigüitat es presenta sempre, per la pròpia natura de les llengües, en el sentit que els significats virtuals dels mots només prenen valors específics, en rigor irrepetibles, quan compareixen dintre de determinades expressions. En part ho reflecteixen els diccionaris monolingües convencionals en distingir entre 'entrades' i 'accepcions'.

Això es posa especialment de relleu en la TA, l'activitat sens dubte més completa del PLN pel fet d'haver de fer conjuminar estructures de dues llengües, si més no. També, en els diccionaris convencionals bilingües cada mot de la llengua d'origen consta almenys d'una entrada lèxica amb un conjunt d'accepcions que equivalen a altres tantes entrades de la llengua de destinació. Per evitar que l'expressió bíblica en llatí *Spiritus quidem promptus est caro vero infirma* es verteixi com a *El licor és al punt, però el bistec està fofo* (en lloc de *L'esperit està disposat, però la carn és feble*) cal establir certament les condicions per les quals les entrades lèxiques, els mots, del text d'origen, passant per les accepcions, donin lloc a les entrades lèxiques adients del text de destinació.[10] Això només es pot aconseguir obrint inevitablement la via a l'anàlisi sintàctica i és en aquest sentit com s'estableix la relació intrínseca entre el diccionari i els mecanismes d'anàlisi, i síntesi sintàctica, els anomenats 'parsers' i generadors en l'àmbit del PLN.[11] A partir d'aquí comencen les diferències.

4.2 Ara bé, en comptes de penetrar en una casuística sobre els aspectes diferencials dels formalismes –que omplirien una pila de tractats–, optarem per presentar un panorama general tot adoptant, a més a més, la perspectiva més favorable de la lexicografia i la TA. La majoria d'analitzadors sintàctics actuals, entre ells els que hem citat a 1.2, tenen en comú que procedeixen de models coneguts per 'gramàtiques d'estats finits', especialment adequats per a la implementació computacional en forma de les anomenades 'xarxes de transició augmentada' (*ATN, Augmented Transition Nets*) impulsades sobretot per W. A. Woods cap a 1970. En desenvolupaments ja clàssics, a més de resoldre diversos problemes de procediment sintàctic (en especial, la recuperació d'informació cap enrere, la supressió de transformacions i l'aplicació sistemàtica de la unificació), havien anat incorporant una diversitat d'aspectes lèxics que analitzarem tot seguit.[12]

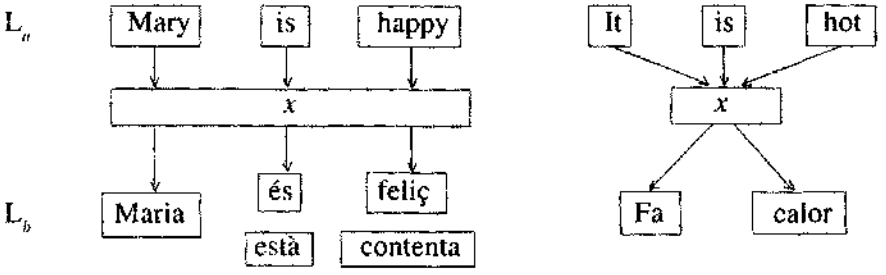
Per això cal aprofitar la distinció, de gran abast teòric i empíric en la TA, que es fa entre la 'transferència' i la 'interlingua'. Considerem-la a través dels següents exemples. Si comparem la traducció de l'anglès al català d'expressions com ara *Mary is happy* i *It's hot*, convindrem que ocupen gairebé posicions extremes en l'escala de la literalitat i la no literalitat. Gràficament ho podem representar així:

(10) L'exemple es refereix a una coneguda facècia que ridiculitzava les primeres experiències en TA mot a mot a partir d'aquell passatge, extret de la *Vulgata*. El mot *spiritus* fou traduït a l'anglès per l'equivalent a "licor", en comptes d'"esperit", i *caro* ho va ser com a "vianda", en lloc de "carn [humana]". El presumpte sistema va escollir accepcions possibles que un traductor humà segurament ni hauria advertit per l'oposició que el propi text estableix entre l'esperit i el cos.

(11) El 'parser', o analitzador sintàctic, és una composició terminològica confegida a partir del llatí *pars orationis* i s'aplica a l'operació de determinar computacionalment les derivacions d'una oració d'acord amb uns principis gramaticals. Més que una excepcional condescendència cap al llatí i als hàbits terminològics de la tradició, el barbarisme *parser* sembla una proclamació de l'actual hegemonia de l'anglès i d'altres tics concomitants (ús d'acrònims, etc.). Quant als termes 'generador', 'generació', provenen de la matemàtica i més en especial de N. Chomsky (vegeu 4.3).

(12) El procediment de la 'unificació', inicialment adduït per M. Kay, és considerat una aportació cabdal en el desenvolupament de formalismes computacionals. Consta bàsicament d'un mecanisme construccional que compara els trets gramaticals de les estructures (arbòries) locals i n'arreplega –unifica– els compatibles per tal de bastir estructures cada vegada més grans formalment definides (vegeu-ne un exemple a 4.3.1).

(9)



En general tota traducció comporta 1) la interpretació d'una estructura d'entrada en una llengua L_a i 2) la seva conversió en una estructura de sortida en una llengua L_b que tingui la mateixa interpretació. (Algú ho ha expressat sintèticament dient que es tradueixen interpretacions i no mots.) En l'esquema representem aquest doble processament amb un rectangle intermedi amb una *x* al mig. En el primer cas de (9) sembla que el processament manté l'estructura gramatical i lèxica del text inicial, tot i que la interpretació queda oberta a dues possibles traduccions en el text de destinació. En el segon cas, en canvi, la interpretació comporta una reorganització total d'estructura gramatical i lèxica. Del pronom *it*, la còpula *is* i l'adjectiu *hot* del text inicial no en queda cap rastre en el pro-verb *fa* i el nom *calor* del text de destinació. Si, malgrat això, la interpretació d'ambdós textos coincideix, caldrà convenir que el processament ha passat per una representació intermèdia neutra o, en qualsevol cas, diferent de les estructures gramaticals de tots dos extrems.

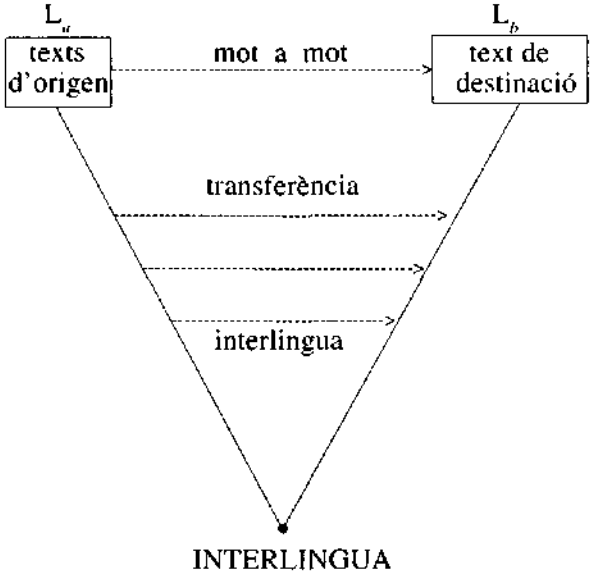
De quina naturalesa és aquesta representació? Molts filòsofs i lingüistes teòrics hi han dedicat grans esforços en les discussions seculars sobre els universals lògics i lingüístics. És la 'gramàtica universal' dels generativistes i la 'interlingua' de la TA, a la qual dóna un innegable suport l'evidència psicolingüística quotidiana dels traductors humans que verteixen sense cap restricció expressions entre llengües il·limitadament allunyades en disseny tipològic. Durant un temps es va pensar que el procediment per excel·lència de la TA consistiria a crear 1) una representació universal, una interlingua, per a tots els continguts lingüístics possibles i 2) les vies d'accés per a cada llengua natural. És, notem-ho bé, la hipòtesi justament contrària a la TA mot a mot. Encara que podia semblar que això complicava l'operació habitual de traduir, en interposar dues passes noves, s'entenia que el model era màximament 'realista' comparant-ho amb l'activitat del parlant humà. El mal és que ningú no ha aconseguit mai ni tan sols albirar de debò com poden ser cap d'aquelles dues fites...

La idea ha estat després replantejada en termes empírics (i amb canvis terminològics, com veurem), a partir d'una altra evidència intuïtiva i estadística igualment incontrovertible: l'afinitat tipològica entre dues o més llengües és directament proporcional al nombre de possibles traduccions literals o quasi-literals correctes entre textos respectius. En rigor, la similitud de les llengües és precisament això. Realistes

o no, en el sentit d'abans, el cert és que les traduccions literals són molt més barates, des del punt de vista procedural, que no pas les no literals. En canvi, les traduccions del segon cas de més amunt junt amb modismes, frases fetes, etc.– requereixen un esforç computacional màxim i molt sovint encara insuficient.[13]

En la mesura, doncs, en què la TA no és profitosa mot a mot ni accessible en la interlingua (en el sentit que hem descrit), s'ha optat per algunes solucions intermèdies, que els especialistes han denominat genèricament de 'transferència'. Gràficament:

(10)



Per anar del text d'origen en la llengua L_a al text de destinació en la llengua L_b es pot prendre la drecera, en la traducció mot a mot, o seguir el trajecte més remot, el que passa per la INTERLINGUA. La distància recorreguda queda compensada amb escreix pel fet que en la drecera les llengües implicades són màximament diferents, mentre que en la INTERLINGUA són una mateixa cosa.

Però no és aquesta l'única consideració important aquí. Com hem vist a 3.1, els sistemes de PLN que actuen sobre textos tancats, és a dir d'abast temàtic limitat i flux de la informació controlat, tenen la característica general d'atribuir valors fixos als elements lèxics utilitzats. I en la mesura en què això és vàlid, el procediment mot a mot o de correspondència de patrons (3.2) és justament l'indicat. Al seu torn, els sistemes que actuen sobre textos oberts, és a dir quan hi ha implicacions contextuals decisives en la determinació del valors lèxics, cal habilitar algun procediment construccional interpretatiu. I en la mesura en què se n'habiliten, el trajecte es fa certament més llarg i complex, però en acostar-se les representacions obtingudes de les llengües implicades la traducció és paradoxalment més directa i acurada. L'esquema (10) assenyala diverses possibilitats en aquests trajectes indirectes, unes més 'profundes' que altres. De menys

(13) Tampoc no han faltat alternatives a favor de llengües pivot preexistents considerades 'transparents' per la seva regularitat. Fa anys que I. Guzmán de Rojas, un informàtic bolivià, ve proposant l'adopció de l'aimarà, llengua ameríndia dotada, entre altres característiques, d'un sistema de sufixos òptim per a la lògica trivalent. L'empresa holandesa DLT utilitza l'esperanto per al seu sistema de traducció, que ja es comercialitza. Evidentment, són només aproximacions relatives a la noció d'interlingua que, com a molt, tenen un abast limitat a certs fenòmens i/o a un grup restringit de llengües naturals.

a més profunditat, en general l'escala estableix anàlisis successives sovint subdividides d'estructura morfològica, sintàctica, semàntica i lògico-conceptual.

Aquestes diferències no són tampoc aquí estrictament quantitatives, sinó que radiquen precisament en la complexitat del tractament lexicològic. Resumint i simplificant (passant per alt tota interferència): abans del nivell sintàctic, es treballa sobre mots o entrades lexicogràfiques predeterminades. Entre els nivells morfològic i sintàctic, es treballa sobre equivalents d'accepcions, és a dir subentrades lexicogràfiques en el sentit que precisarem a 4.3.2. A partir del nivell semàntic, es treballa sobre trets sèmics més abstractes.

Aquí hi ha, com dèiem, una impropietat terminològica, en el sentit que s'acostuma a aplicar la denominació d''interlingüístic' a tot sistema que actua més enllà dels nivells estrictament sintàctics, per molt lluny que estiguin encara d'arribar a la INTERLINGUA tal com l'hem definida abans.

Adoptant una perspectiva lexicològica, analitzarem per separat les principals característiques dels sistemes així caracteritzats, d'accepcions i de trets sèmics oracionals, tot recordant que en TA la complexitat del procediment ve molt determinada per la distància tipològica que separa les llengües implicades. Altrament dit, si les llengües incurses són prou semblants –típicament quan pertanyen a una mateixa família genètica, com portuguès i italià, neerlandès i alemany o mandarí i cantonès–, es podrà utilitzar un sistema de transferència. Per a llengües prou allunyades (entre japonès i hausa o basc, o entre elles i qualsevol de les anteriors), el sistema de transferència serà insuficient o, com a poc, extremament complex. I és a partir d'un cert grau de complexitat –especialment en el pas de solucions lèxico-sintàctiques a solucions semàntiques– quan se sol parlar d'interlingua.

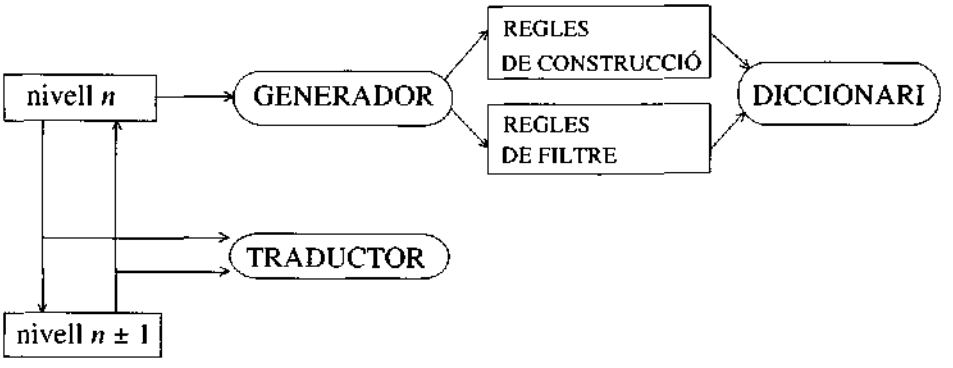
Nosaltres no sols evitarem aquesta confusió terminològica, sinó que aprofitarem la distinció entre 'interlingua' i 'INTERLINGUA' per a al·ludir, respectivament, al processament lingüístic semàntic i conceptual i establir així el límit entre el PLN i la IA. Al primer li dedicarem el pròxims quatre subapartats, i a la segona, un últim apartat sencer.

Un parell d'observacions encara, tornant a l'esquema (10). La part esquerra, la diagonal que va del text d'origen fins a un cert punt en direcció a la INTERLINGUA, indica el procés, denominat d'anàlisi, pel qual el text en qüestió queda interpretat o desambiguat, tècnicament convertit en una certa representació formal o canònica (morfològica, sintàctica...) del seu contingut (vegeu-ne una aplicació a 4.3.2 (21)). La diagonal dreta indica el procés invers, denominat de síntesi o generació, pel qual la representació formal del contingut provinent de l'anàlisi, un cop procesada en alguna mesura pel nexe d'enllaç, de transferència o interlingua, rep una materialització lèxica i és convertida en el text de destinació. Són, respectivament, els dominis del "parser" i del generador. Estrictament parlant, els sistemes de TA es caracteritzen només pel nexe que enllaça l'anàlisi amb la generació de textos, cadascun en llengües diferents. La resta d'elements, és a dir l'anàlisi i la generació són per ells mateixos propis del

PLN en general (per bé que en la TA s'orienten específicament cap al nexe d'enllaç o provenen d'ell, també respectivament).

4.3.1 Els sistemes que treballen sobre accepcions són, en realitat, mecanismes lexicogràfics no trivials que interactuen modularment amb gramàtiques. El "parser" típic d'un sistema així consta d'una sèrie ordenada de *n* nivells o mòduls que confereixen una o més representacions formals a partir de les representacions del nivell anterior o d'un text d'origen ja preeditat (del tipus que veurem a 4.3.2 (21)).[14] Cada nivell està constituït per una gramàtica, que a vegades s'anomena (no massa feliçment) 'generador', composta d'una col·lecció de regles de diversos tipus, i un diccionari associat. Entre els nivells hi ha, com és lògic, nexes interiors, que consten igualment d'una col·lecció de regles, que anomenarem 'traductor', per projectar o transferir les estructures d'un nivell al següent en la direcció que convingui a l'anàlisi o a la generació. Gràficament seria:

(11)



Examinem més de prop el funcionament de les gramàtiques i els diccionaris en un formalisme d'unificació amb un exemple força sumari. Suposem que el text d'origen sigui una oració com *La cartera porta un sobre*. La primera operació verifica si les formes de mot són reconegudes pel diccionari morfològic. Diguem que reconeix efectivament cinc formes, a les quals assigna categories sintàctiques corresponents (aquí simplificades):

(12)
(1) [la] ..... { *Pronom* [3 fs] / *Determinat* [3 fs]
(2) [cartera] ..... *Nom* [3 fs]
(3) [porta] ..... { *Nom* [3 fs] / *Verb* [3 spr]
(4) [un] ..... { *Pronom* [3 ms] / *Determinat* [3 ms]
(5) [sobre] ..... { *Nom* [3 ms] / *Preposició*

(14) Per 'preeditar' s'entén la indicació manual en un text de les expressions que cal no processar –noms propis, xifres, fórmules químiques...– o que han de rebre alguna consideració especial –cursiva, nota al peu...–. Es correspon amb la 'postedició', que comprèn el conjunt d'operacions igualment manuals que s'introdueixen en el text que resulta de la generació per tal de fer-lo més correcte o intel·ligible als usuaris humans.

Pel que fa als trets morfològics, totes les formes són singulars, les tres primeres femenines i les altres dues masculines. I pel resultat de l'escorcoll amb els cinc mots es poden produir setze combinacions i, per tant, altres tantes interpretacions morfològiques possibles:

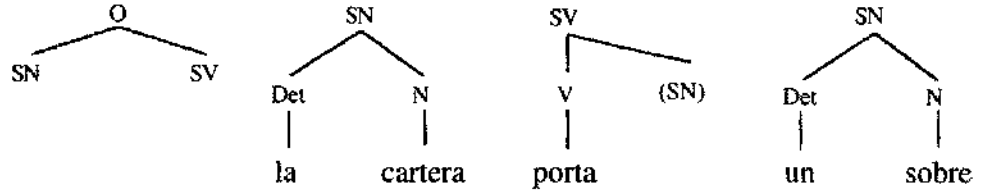
(13) (1) Pronom Nom Nom Pronom Nom

...

(16) Determinant Nom Verb Determinant Preposició

En l'anàlisi sintàctica 'superficial', les regles de construcció componen totes les alternatives combinatòries i les de filtre consulten primer quines de les possibles seqüències són gramaticals o no i després quines, entre les gramaticals, són prioritàries. La primera comprovació consisteix a comparar les cadenes potencials amb un catàleg d'estructures gramaticals oracionals i sintagmàtiques possibles. En el nostre exemple, la seqüència (1) és clarament agramatical, ja que només consta de sintagmes nominals i no correspon a cap estrucura oracional catalogada. Com que cal almenys un sintagma verbal perquè hi hagi una oració, s'ha d'assignar la categoria {verb} a la forma *porta* (l'única que l'accepta, segons l'anàlisi anterior) i descartar les altres combinacions. La (16) també és agramatical, tot i que duu un verb, perquè l'assignació de la categoria {preposició} a *sobre* requereix un sintagma nominal regit, que falta. Al seu torn, cal donar prioritat a l'assignació de categoria {determinant} (i no {pronom}) a les formes *la* i *un* perquè la seva unificació de trets amb els noms *cartera* i *sobre*, respectivament, produeix sintagmes nominals ben formats i més complets, en el sentit que componen una estructura global més compacta, sense formes soltes. De tot això en resulta una estructura oracional consolidada i tres de sintagmàtiques locals, dues nominals i una verbal:[15]

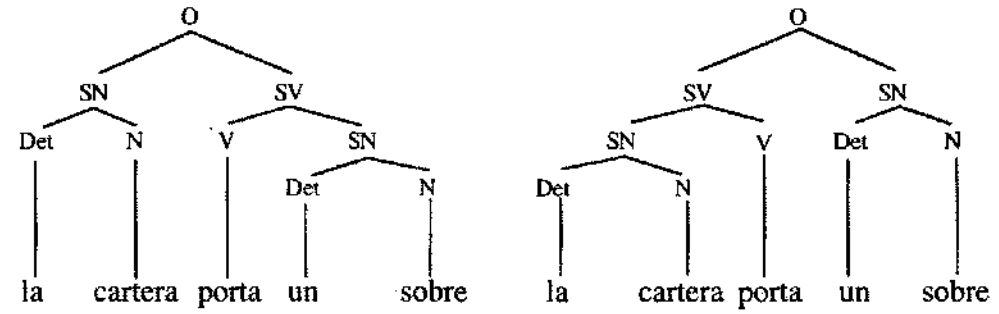
(14)



Comptant que el sintagma verbal tingui o no un sintagma nominal regit (dit d'altra manera, que el verb *portar* funcioni transitivament o no), hi ha quatre possibles estructures oracionals, segons que el primer sintagma nominal sigui ocupat per *la cartera* o per *un sobre* i que *porta* sigui transitiu o no. Les regles de construcció corresponents compondran totes aquestes possibilitats. Les de filtre establiran una prioritat a favor de l'ús transitiu de *porta*, perquè és l'únic que deixa ocupats tots els nusos de l'arbre general. Això produeix, encara, dues possibilitats oracionals:

(15) Els diagrames que segueixen a títol purament il·lustratiu ometen tota formulació de sintaxi d'X amb barres.

(15)



La consulta al diccionari d'aquest nivell d'anàlisi fornirà la subcategorització de *portar*, és a dir el nombre i l'estructura construccional dels seus arguments i els trets dels nuclis corresponents. Una cosa així com:

(16) PORTAR categoria verb; arg1 nom[+A]; arg2 nom[_A]; ...

L'entrada estableix que *portar* és verb que es construeix amb un subjecte nominal (argument 1) marcat amb el tret [+animat] i amb un objecte també nominal (argument 2) neutre per a l'animació.[16] Aquesta dada pot servir per establir, per via unificacional, la funció dels dos sintagmes nominals candidats si es dóna el cas que només un pot dur el tret [+animat]. Consultant les entrades dels nuclis respectius, trobem efectivament:

(17) (a) CARTERA_1 categoria nom; [+A, +H]
(b) CARTERA_2 categoria nom; [-A]

(18) (a) SOBRE categoria nom; [-A]

(aquí ja no compta l'opció de *sobre* com a preposició). L'entrada (17) (a) (equivalent, en certa manera, a una accepció del diccionari convencional) es refereix òbviament a la persona del sexe femení que treballa de repartidora a correus; la segona, (17) (b) als contenidors de mà per bitllets o documents, per anar a col·legi, etcètera. Amb això, arribem finalment a consolidar una estructura oracional única que coincideix amb la primera que hem vist a (15).[17]

4.3.2 Fem una nova aproximació. Comparem ara una entrada lèxica real, la del verb *elevar*, en un diccionari convencional:

(16) La subcategorització real de *portar* inclouria altres arguments (de destinació, etc.), que aquí ometem per simplificar.

(17) Posem que recapitulem algunes qüestions. El tret [+animat] es refereix a la capacitat d'un ésser de moure's per ell mateix i és característica pròpia dels 'animals' (humans o no). A més a més, la nostra gramàtica estableix que *portar* duu subjecte animat i descarta, per tant, no sols la combinació equivalent a **Un sobre porta la cartera*, sinó també **La cartera porta el sobre* entenent *cartera* com a [-animat]. Això obre un seguit de temes teòrics que miraré de resumir tan de pressa com pugui. En primer lloc, no hi ha d'haver cap impediment per sostenir la agramaticalitat de l'última construcció considerada en un registre lingüístic bàsic (on, en tot cas, s'hauria de canviar de verb i dir, per exemple, *La cartera conté un sobre* per a entendre *cartera* en sentit [-animat]). Computacionalment parlant, si més no, és indispensable establir amb claredat aquests tipus de restriccions, que marquen els límits de la gramaticalitat, encara que la gent pugui ampliar-los en altres registres (literaris, humorístics...). D'aquí a una mica ens referirem a una qüestió connexa, la caracterització lexicogràfica dels textos a propòsit de la terminologia. En segon lloc, noteu que aquesta disquisició sobre el caràcter animat o inanimat de *cartera* queda inèdita en el resultat final que forneix el sistema. Si sabem que el sistema entén *cartera* com a [+animat] només és perquè n'hem seguit el procés intern. I si haguéssim de traduir aquest resultat a l'anglès, al francès i a l'alemany, trobaríem en aquestes llengües sengles distincions lèxiques per al tret [-animat]. Molt per sobre:

elevar *v* **1** ***a*** *tr* Pujar, transportar a un nivell més alt. *Elevar l'aigua d'un pou mitjançant una bomba.* ***b*** *pron* Pujar, anar cap amunt. *Si no fa vent, el fum s'eleva ben dret.* ***c*** *tr fig* Portar a un càrrec, dignitat, grau, etc, superior. *Elevar un príncep al tron.* ***d*** *pron fig S'ha ele-vat a la presidència il·legalment.* ***e*** *tr fig* Posar més amunt, molt amunt. *Elevar els sentiments del poble.* ***f*** *pron fig El seu pensament s'eleva en especulacions filosòfiques.* ***g*** *tr fig* Transportar a un alt grau de contemplació, superior als sentits. *Elevar l'ànima envers Déu.* ***h*** *pron fig S'eleva en contemplacions místiques.* ***i*** *tr* Dirigir un escrit oficial a una autoritat, persona de categoria superior, etc. *Elevar una petició al govern.* **2** ***a*** *tr* Fer més alt, estendre fins a una major alçària. *La fosa de la neu ha elevat el nivell del riu.* ***b*** *pron En aquest barri industrial, el nivell de la pol·lució atmosfèrica s'eleva un 20%.* ***c*** *tr* Tenir a una certa alçària. *L'alzina elevava les seves branques per sobre dels altres arbres.* ***d*** *pron Aquell cim s'eleva molt més que els altres.* ***e*** *tr* Erigir, construir (un temple, un monument, etc) en honor d'algú. ***f*** *pron* Estar situat, existir en un lloc determinat, un edifici d'una notable alçària, una muntanya, etc. *Al costat del riu s'eleva un gran monestir.* **3** *fig* ***a*** *tr* Augmentar en intensitat, grau, quantitat, etc. *La calor eleva el volum dels gasos.* ***b*** *pron Aquesta setmana han elevat els preus dels llegums.* **4 elevar a una potència** *mat* Obtenir la potència d'un nombre. **5 elevar una perpendicular** *geom* Traçar una perpendicular a una recta en un punt determinat.

19. Entrada **elevar** del *Diccionari de la Llengua Catalana* Enciclopèdia Catalana, 1982.[18]

amb les entrades lèxiques que el mateix verb té en el sistema EUROTRA:[19]

elevar_1={cat=v,e_lu=elevar,string=elev,thcat=string,infl=stem,pres_s1=t1,pres_s23p3=t1,pres_p12_inf=t1,impf_type=t1,past_s1=t1,past=s2p12=t1,past_s3=t1,past_p3=t1,futcond_type=t1,pressubj_s123p3=t1,pressubj_p12=t1,impfsubj_type=t1,ger_type=t1,part_type=t1,imper_s2=t1,imper_s3_p3=t1,imper_p1=t1,imper_p2=t1,pron=nil,pronlu=nil}.

elevar_1={cat=v,e_lu=elevar,recform=nil,recpart=nil,fpformcomp=nil,fpformcomp2=nil,fp1type=dest,fp2type=nil,fatr=no,fpas=yes,pron=no}.

elevar_1={cat=v,e_lu=elevar,ersfr=acnpronpas,e_pformarg1=nil,e_pformarg2=nil,e_pformarg3=nil,e_pformarg4=nil,p1type=dest,p2type=nil,control=nil}.

elevar_1={cat=v,e_lu=elevar,e_isrno_'1',eisframe=arg1_2_goal,e_pformarg1=nil,e_pformarg2=nil,e_pformarg3=nil,e_pformarg4=nil,p1type=dest,p2type=nil,semarg1=conc,semarg2=ent,semarg3=ent,semarg4=nil,e_vtype=main,vfeat=nstat,attype=nil,instrumental=yes,term='0',erg=yes}.

20. Entrada **elevar** als diccionaris d'EUROTRA[20]

que corresponen respectivament a quatre nivells de representació o 'estructures', ordenades aquí per a l'anàlisi:

(21) (1) **estructura morfològica**, amb indicació precisa del model pertinent de flexió
(2) **estructura configuracional**, anàlisi sintàctica superficial a base de possibles agrupacions de constituents ordenats tal com estan en el text

	[+animat]	[-animat]
català	cartera	cartera
francès	facteuse	portefeuille porte-documents pochette cartable
anglès	postwoman	wallet briefcase portfolio pocketbook
alemany	Briefträgerin	Brieftasche Aktentasche Bestand Schulmappe

La traducció des del català donaria solucions lèxiques úniques en tots els casos, però, tal com està, resultaria inevitablement ambigua si *cartera* quedés caracteritzat només amb el tret [-animat].

(18) Ben mirat, no sempre és fàcil endevinar el raonament lexicològic dels autors, atesa la barreja de sentits i de propietats sintàctiques dels exemples.

(19) EUROTRA és un sistema de traducció automàtica de disseny avançat auspiciat per la Comunitat Europea i desenvolupat entre 1983 i 1992. Comprèn alhora nou llengües, totes pertanyents a la branca occidental indoeuropea: quatre de germàniques (danès, alemany, neerlandès i anglès), quatre de romàniques (portuguès, espanyol, francès i italià) i el grec. Pel seu abast científic, lingüístic i econòmic ha estat, durant aquell període, el més extens entre els seus coetanis. Després de 1992 se segueix desenvolupant amb altres denominacions, formalismes (ALEP) i estructura organitzativa.

(20) En rigor es refereix al verb castellà *elevar* en una accepció (l'entrada '1') equivalent a **1-*i*** de (19).

(3) **estructura relacional**, anàlisi sintàctica amb assumpció (i eliminació visual) d'auxiliars i altres partícules, assignació de funcions oracionals (subjecte, objecte...) i canonització de l'ordre de mots[21]

(4) **estructura d'interfície**, assumpció i eliminació de determinants, preposicions i trets verbals com temps i aspecte, assignació de 'rols' d'estructura profunda (nuclis, arguments...)

El sistema operacional d'EUROTRA és del tipus anomenat de transferència (atesa la relativa similitud tipològica de les nou llengües amb què treballa) i està construït de tal manera que la transferència al mòdul de generació en l'altra llengua sigui el més simple possible i, si pot ser, inexistent. Altrament dit, l'estructura d'interfície que culmina l'anàlisi del text d'origen, escrit en la llengua L_a, tendeix a ser tot l'idèntica que es pot a l'estructura d'interfície amb què comença la generació del text de destinació, escrit en la llengua L_b. I quan diem idèntica, volem dir que, en aquell punt del procés, només cal substituir les entrades lèxiques de L_a per les corresponents entrades lèxiques de L_b en la mateixa representació arbòria.[22] Això és, tot plegat, el que justifica tantes i tan costoses operacions i el que ens duu amb naturalitat a la definició acurada d''accepció' dintre del PLN.

Acabem de veure a (20) que el diccionari és una col·lecció d'entrades, cadascuna d'elles caracteritzada per un conjunt de trets, el qual no es limita, com en l'exemple de *cartera*, a [±animat], sinó que en comprèn molts altres, segons el sistema, com ara [±humà], [±part_del_cos], [±vegetal], [±comestible], [±objecte], [±massa], [±líquid], [±continent], [±moble], [±localització], [±instrument], [±institució], etc., etc. Al seu torn, la gramàtica és un mecanisme que 1) subministra un catàleg d'estructures arbòries possibles, 2) localitza en el diccionari els mots del text d'origen i els assigna els trets de subcategorització que hi troba, 3) compon totes les estructures arbòries que pot amb els nusos ocupats pels mots del text definits per sengles conjunts de trets de subcategorització, i 4) mitjançant la unificació, acobla els conjunts de trets compatibles i filtra successivament les estructures arbòries obtingudes en virtut de la seva coherència amb les estructures catalogades. Aleshores, cada entrada del diccionari, delimitada per un subconjunt de trets, constitueix un possible nus d'una estructura arbòria o, si es vol, un tipus definit de comportament composicional formalment recognoscible pel formalisme d'unificació. Comparant l'entrada d'aquest diccionari amb la d'un de convencional, com el de (19), s'hi troba una correspondència, almenys tendencial, de la primera amb aquelles (sub)accepcions de la segona que presenten alhora propietats semàntiques i sintàctiques ben definides. Dintre del que els constructors del sistema han decidit o han pogut delimitar, es pot dir que tantes accepcions així com hi hagi, tantes entrades hi haurà en el diccionari computacional. Cal advertir finalment que en PLN la noció d'accepció equival a la d'interpretació, ambdues en els sentits descrits.

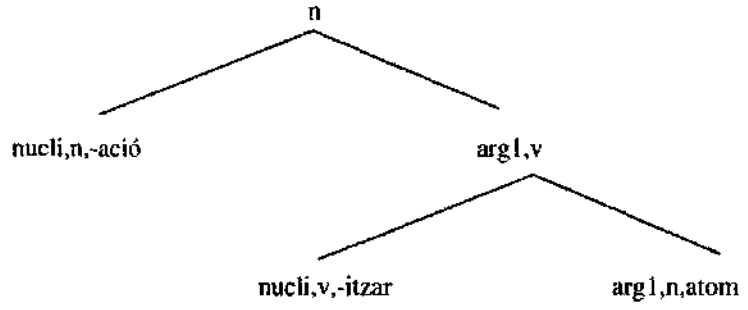
4.3.3 A part d'entrades generals com la que hem considerat a (20), el diccionari

(21) Per assumpció (tècnicament 'elevació') de trets o elements s'entén la incorporació del seu contingut en la matriu estructural que no es visualitza en el diagrama, però que acompanya els arbres locals o generals. Els continguts així assumits poden ser visualitzats de nou en la representació simètrica de la generació, si la llengua L_b els lexicalitza d'alguna forma. I per 'canonització d'ordre' s'entén l'adaptació que el sistema fa a una seqüència fixa d'elements en les estructures locals: primer el nucli i després els arguments, igualment ordenats.

(22) En aquest cas, la transferència consta únicament d'un diccionari de correspondències lèxiques. En circumstàncies menys favorables conté també regles de reajustament.

computacional en conté d'altres que tampoc no coincideixen amb les rèpliques convencionals. Es tracta, d'una banda, de morfemes, clítics i partícules derivacionals que entren en la composició ortogràfica (o fònica, si s'escau) de les entrades generals d'acord amb les corresponents descripcions morfològiques. Mitjançant especificacions del tipus que hem vist a la primera subentrada de (20), corresponent a l'estructura morfològica, es poden descompondre i recompondre les formes superficials dels mots en procés. Aquesta funció la realitza el mòdul fonològic, sovint constituït per un conjunt de regles de dependència no lineals que estableixen no sols l'ortografia (o la pronúncia, en el seu cas) correcta dels mots, sinó també les possibles derivacions composicionals. Les regles fonològiques solen aprofitar l'estructura i la informació que forneix el mòdul sintàctic superficial fins al punt que n'identifiquen els límits en un grau considerable. Així, per exemple, ho reflecteix la composició d'*atomització* i *atomitzar* a partir del nucli nominal *àtom*:

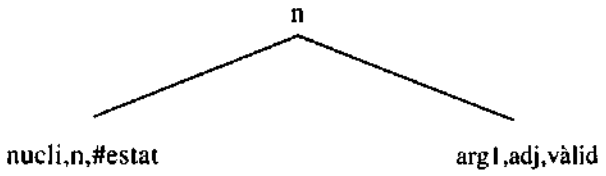
(22)



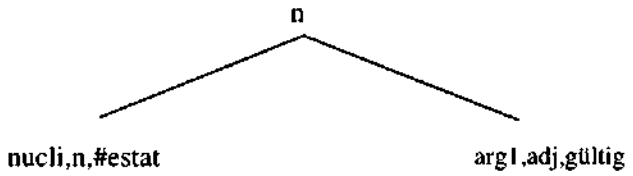
Començant per baix a la dreta, la seqüència *atom*, que pertany a la categoria nom, dóna lloc com a argument a una composició verbal amb un nucli que aporta el sufix *-itzar* i produeix *atomitzar*. Aquesta seqüència derivada intervé al seu torn com a argument en una altra derivació (deverbal) subsegüent en què, mitjançant el sufix *-ació*, produeix el resultat *atomització* amb categoria nom, gràcies també a la cooperació d'altres regles més generals associades que recomponen els marges dels segments afectats i estableixen els valors vocàlics i consonàntics de cada derivació.

Els mateixos principis de morfologia composicional permeten simplificar moltíssim les entrades dels diccionaris bilingües sempre que hi hagi, és clar, prou analogia en els processos derivacionals, com es desprèn del següent exemple sobre la nominalització de l'adjectiu *vàlid* i el seu equivalent, *gültig*, en alemany, on la indicació de significat "estat" selecciona automàticament el sufix adient:

(23) català *validesa*



alemany *Gültigkeit*



D'altra banda, hi ha els mots compostos, en qualsevol modalitat morfològica (*no_obstant, cop_de_mà, portallapis*...). En sistemes de TA per transferència sobretot entre llengües aglutinants i flexives, els compostos solen necessitar un complex tractament especial, atès que moltes solucions lèxiques a L_a requereixen interpretacions sintàctiques a L_b i viceversa (v. gr. alemany *Einbahnstraße* ↔ català *carrer de direcció única*).

Finalment, queda la terminologia (*aldehid, endogàmia, hipotenusa, reostat*...), en què el tractament és tot altre. La seva estricta equivalència lèxica a través de les llengües afavoreix la correspondència de patrons (3.2), molt més econòmica. Això comporta que els termes tècnics formin diccionaris específics d'utilització preferencial, almenys per dues raons: 1) perquè es tradueixen directament mot a mot sense entrar en el processament general del sistema, i 2) perquè poden compondre terminologies temàticament homogènies al marge del diccionari general. D'aquesta manera, l'eficiència dels grans sistemes de PLN augmenta molt significativament si funcionen amb dos diccionaris, un de general, amb els fenòmens gramaticals i les entrades més freqüents de la llengua, i un altre de temàtic incorporat, que recull només la terminologia especialitzada dels textos en procés.[23] Vet aquí la fesomia bàsica d'un diccionari electrònic en aquest nivell de complexitat.[24]

4.4 Molt sovint la distància tipològica entre les llengües naturals implicades en un sistema de TA és massa gran i no n'hi ha prou amb un sistema de transferència, basat, en definitiva, en un escorcoll lexicogràfic no trivial com el que acabem d'analitzar en els paràgrafs anteriors. El sistema de transferència, tot i que incorpora trets semàntics,

(23) La tipologia textual atenua una de les deficiències que caracteritza els sistemes de PLN enfront del parlant humà: la manca de transportabilitat, és a dir la incapacitat de canviar més o menys radicalment de tema sense modificar res en aparença. Això val també per la cobertura gramatical de les regles. Si, per exemple, es processa un text amb instruccions per al muntatge, manteniment i ús d'un vídeo, un sistema optimitzat negligirà fenòmens gramaticals com ara les construccions de passat, les concessives o les condicionals eventuals i irreals.

(24) El terme 'electrònic' s'aplica a vegades per designar qualsevol diccionari convencional o text classificat que es presenta en suport informàtic (programes residents, en CD-Rom, hipertextos, etc.). Es tracta, evidentment, de productes ben diferents. Altres vegades s'aplica, amb una mica més de propietat, a diccionaris amb entrades simples (arrels, derivacions, desinències...) o compostes (modismes, col·locacions...) caracteritzades per uns quants trets morfològics i sintàctics. Poden marcar (semi)automàticament extensos corpus de referència i servir per un seguit de comeses lexicològiques o gramaticals. Són, en tot cas, diccionaris electrònics de baix nivell, comparats amb els que hem vist.

té un fonament essencialment sintàctic en el sentit que radica en la subcategorització d'entrades lèxiques de diccionari, estructures arbòries i funcions oracionals com subjecte, objecte, etc. La seva eficàcia depèn del percentatge de traduccions literals o quasi-literals com les que hem vist a (9). Però no serveix quan no hi ha literalitat o, si més no, tan bon punt com la subcategorització de les entrades resulta insuficient.

Els sistemes que transcendeixen la sintaxi i creen una nova dimensió composicional s'inscriuen en el mètode d'interlingua, com hem dit. Però tot i així, encara s'hi poden identificar graus de complexitat que vénen a delimitar els camps del PLN i de la IA, segons que el processament romangui dintre d'estructures lingüístiques o les ultrapassi –diferència que abans hem descrit oposant 'interlingua' amb 'INTERLINGUA'–. En realitat, són dos trams successius d'una sola direcció, que va d'una representació semàntica fins a una representació conceptual, ambdues obertes a diverses alternatives teòriques. Examinem-ne les principals –sempre des de la perspectiva lexicogràfica, fins on sigui possible– començant per la representació semàntica a base de valències casuals (*case frames*) inspirades inicialment en l'anomenada 'gramàtica dels casos' de Ch. Fillmore.

Si comparem les següents expressions:

(24) (a) Joan obre la porta amb la clau
(b) Joan obre la porta
(c) La clau obre la porta
(d) S'obre la porta
(e) La porta és oberta per Joan
(f) La porta és oberta per la clau

admetrem sense dificultat que comparteixen molta part del significat tot i que es diferencien per l'estructura sintàctica. El subjecte, per exemple, a (a) i (b) és *Joan*, a (c) *la clau*, i a (d), (e) i (f) *la porta*. Però *Joan, la clau* i *la porta* presenten també altres funcions. I també es veu que (e) és la versió passiva de (b), com (f) l'és de (c), i que, per tant, es pot entendre que signifiquen el mateix i que provenen d'una sola representació més 'abstracta'. Nosaltres entendrem que totes aquestes expressions provenen d'un sol significat i d'una mateixa representació, lògicament semàntica.[25]

En un treball ja clàssic, Fillmore va proposar l'adopció de 'casos' semàntics profunds (ben diferents dels declinacionals: nominatiu, genitiu...) per descriure amb precisió la funció dels elements que intervenen en una acció i la seva relació mútua.[26] Així, el significat comú convingut de totes les expressions de (24) ve a establir que si Joan executa l'acció d'obrir la porta amb una clau, *Joan* rep el cas 'agent', *porta* el cas 'objecte' i *clau* el cas 'instrument'. Fillmore va definir encara més casos, com ara 'locatiu' (*al jardí*), 'benefactiu' (*per al gat*), 'direccional' (*cap a fora*) i d'altres que s'han anat incrementant i refinant amb el temps. Una representació abreujada dels exemples de (24) (a)-(f) anteriors amb valències casuals seria:

(25) El supòsit que prenem en aquesta il·lustració no és pas que les expressions de (24) hagin de tenir una mateixa interpretació –cosa que seria falsa–, sinó a l'inrevés: suposant que Joan obre efectivament la porta amb la clau, entenem que aquesta acció pot ser descrita per totes i qualsevol de les expressions de (24).

(26) «Case for case», a Bach, E. & Harms, R. (comps.) (1968) *Universals in Linguistic Theory*, Holt, Rinehart & Winston, pp. 1-88. En contrast amb la teoria transformacionalista, la contribució de Fillmore va palesar immediatament la seva aplicabilitat computacional.

(25)

(a) [OBRIR
[valència casual
agent: *Joan*
objecte: *porta*
instrument: *clau*
...]
[modals
temps: present
veu: activa]]

(b) [OBRIR
[valència casual
agent: *Joan*
objecte: *porta*
instrument: ϕ
...]
[modals
temps: present
veu: activa]]

(c) [OBRIR
[valència casual
agent: ϕ
objecte: *porta*
instrument: *clau*
...]
[modals
temps: present
veu: activa]]

(d) [OBRIR
[valència casual
agent: ϕ
objecte: *porta*
instrument: ϕ
...]
[modals
temps: present
veu: mitja]]

(e) [OBRIR
[valència casual
agent: *Joan*
objecte: *porta*
instrument: ϕ
...]
[modals
temps: present
veu: passiva]]

(f) [OBRIR
[valència casual
agent: ϕ
objecte: *porta*
instrument: *clau*
...]
[modals
temps: present
veu: passiva]]

Noteu que això permet especificar amb molta exactitud les diferències entre les distintes versions sintàctiques de (24) (a)-(f) com a mers canvis de perspectiva que l'acció pren des dels elements incursos i com una elisió d'alguns –possiblement per ignorància o negligència del parlant real–. D'aquí ve que les funcions sintàctiques superficials es permutin entre els elements i que en posició de subjecte hi pugui comparèixer l'agent de l'acció (*Joan*), l'objecte (*la porta*) o l'instrument (*la clau*). Naturalment, tot això té la seva transcendència tant teòrica com metodològica.

Per una banda, aquest procediment comporta augmentar un nivell de representació en el sistema de processament tant analític com generatiu. I encara que la representació mateixa sigui simple, el conjunt de regles requerit per connectar les funcions sintàctiques superficials amb les valències casuals és força complex. En l'exemple d'abans no és gaire difícil establir les valències d'*obrir* si en una certa estructura sintàctica apareix acompanyat de *Joan, porta* i *clau*, gràcies als trets constitutius ben diversificats de les respectives entrades lèxiques (és impensable qualsevol interpretació de tipus **La porta obre la clau amb Joan* i coses per l'estil). Habitualment, però, les estructures són alhora alambinades, ambigües i estructuralment incompletes, i es fa imprescindible harmonitzar molta informació sobre l'ordre superficial, sobre dades aparegudes en altres estructures més o menys allunyades i sobre la possible interacció dels referents implicats en el món que es descriu.

Per la banda contrària, aquest sistema forneix una plantilla molt simple i atractiva per als esquemes verbals primaris, que pot servir, en principi, per gairebé tots els verbs en totes les combinacions oracionals de superfície. L'estructura de valències d'*obrir* coincideix amb la de nombrosíssims verbs que descriuen l'acció d'un agent sobre un objecte mitjançant un instrument: *construir, destruir, pintar, escriure, arreglar, guarir...* Sovint la diferència es redueix al nombre de valències i a l'obligatorietat/opcionalitat/impossibilitat de construir-se amb tals o tals altres valències; v. gr. *preferir* requereix obligatòriament un objecte, *menjar* pot tenir-lo o no i *despertar* no pot tenir-lo. Comproveu-ho.

En conjunt, el profit extret a aquestes estratègies ha estat d'allò més remarcable. Comptant només les valències obligatòries (elidides o no, en la parla), no sols es pot bastir una classificació entre els verbs de valència zero (*ploure*), u (*tossir*), dos (*voler*), tres (*regalar*) o quatre (*portar*), sinó que també es poden identificar procediments lèxico-sintàctics per passar d'una classe de verbs a la següent i viceversa.[27] Tal és el cas de la 'causativitat', una mena de recursió d'agentivitat amb el pro-verb *fer* segons una fórmula de tipus [X fa que *Y*] (on X és un agent i *Y* una construcció verbal de valència *n*). Per exemple:

(26) (a) [[Joan] es mor_1] + [caus]: [Pere fa que [[Joan] es $mori_1$] ⇒ [Pere] $mata_2$ [Joan] + [caus]: [Carles fa que [Pere] $mati_2$ [Joan]/[Carles] *fa matar* [Joan] + [caus]: [Albert fa que [Carles faci que [Pere] $mati_2$ [Joan]/[Albert] *fa fer matar* [Joan], etc.

(b) [[El gos] $passeja_1$] + [caus]: [Pere fa que [[el gos] $passegi_1$] ⇒ [Pere] $passeja_2$ [el gos]

(c) [[Joan] $té_2$ [un llibre]] + [caus]: [Pere fa que [[Joan] $tingui_2$ [un llibre]] ⇒ [Pere] $dóna_3$ [un llibre] [a Joan]

d'on es desprèn que *matar* és el causatiu lèxic de *morir* (tal com *fer matar* és el causatiu analític de *matar*, *fer fer matar* l'és de *fer matar*, etc.), que $passejar_2$ és el causatiu lèxic de $passejar_1$ (però sense lexicalització) i que *donar* l'és de *tenir*, al marge d'altres propietats sintàctiques (com ara que *tenir* no sigui pròpiament transitiu i *donar* sí).[28]

Explotant l'enorme avantatge metodològic d'establir relacions d'aquest tipus entre esquemes de valències verbals, R. C. Schank i d'altres que treballaven en línies més o menys similars van fer veure que es podia reduir molt més encara el nombre de representacions bàsiques i que si aquestes eren prou abstractes es transcendien les estructures pròpiament gramaticals de les llengües naturals i s'arribava a una configuració veritablement INTERLINGÜÍSTICA dels significats, a la qual tendirien totes les expressions reals dels parlants.

Aquesta aproximació conceptual de les llengües ens introdueix en el domini de la IA.

(27) S'ha discutit força sobre el nombre màxim de valències obligatòries (més o menys equivalents als arguments de 4.3.1) i sembla que hi ha raons per deixar-lo en quatre: compareu [ϕ] $plou_0$, [Joan] tus_1, [Joan] vol_2 [una moto], [Joan] $regala_3$ [fruita] [a la Maria], [Joan] $porta_4$ [llegums] [de l'hort] [a casa].

(28) Un altre dimensió transcendental és que l'examen aprofundit entre les valències dintre els esquemes verbals ha revelat característiques sèmiques –distincions entre estats, events, processos tèl·lics o no, etc.– que interactuen d'una manera molt precisa i prou complexa amb categories tan diverses com la *Aktionsart*, el temps i l'aspecte.

5. LEXICOGRAFIA I FORMALISMES EN IA

5.1 La idea de Schank era una versió reduccionista de les valències casuals, anomenada 'dependència conceptual', especialment dissenyada per representar esquemes prototípics d'acció. Un exemple clàssic lliga expressions com a) *Joan va donar un llibre a Maria* i b) *Maria va agafar un llibre a Joan* a través d'una sola relació conceptual de transferència batejada amb el nom de 'ATRANS' (*Abstract TRANSfer*). Formalment, i passant per alt els trets irrellevants:

(27)	(a)	[ATRANS	(b)	[ATRANS
		relació: *possessió*		relació: *possessió*
		actor: *Joan*		actor: *Maria*
		objecte: *llibre*		objecte: *llibre*
		origen: *Joan*		origen: *Joan*
		recipient: *Maria*		recipient: *Maria*

A més a més d'establir una relació de possessió, la transferència pot ser de localització, de pertinença, etc. Definida en general, la transferència es troba en la base significativa d'una bona multitud de verbs (o categories similars) en totes les llengües naturals, en el sentit que captura almenys les inferències veritatives essencials de les respectives proposicions. La dependència conceptual queda reduïda a una col·lecció d'accions primitives com ara:

(28)
- (a) ATRANS, transferència abstracta
- (b) PTRANS, canvi físic de localització
- (c) MTRANS, transferència mental d'informació
- (d) MBUILD, creació d'una nova idea o d'una conclusió a partir d'una informació
- (e) INGEST, introducció d'una substància dins el cos
- (f) PROPEL, aplicació d'una força sobre un objecte
- (g) SPEAK, producció de qualsevol mena de so

etc.

cadascuna acompanyada dels esquemes pertinents per copsar-ne subaccions igualment primitives, és a dir universals, sota el principi que hi ha efectivament moltes relacions lògiques –de tipus tot/part, hiperonímia/hiponímia, etc.– independents de qualsevol circumstància lingüística, cultural o històrica.

Examinem-ne algunes conclusions des del punt de vista de la lexicologia i de la confecció de diccionaris. En primer lloc, cal reconèixer que el procediment de la dependència conceptual permet paradoxalment recuperar valors fixos en textos oberts, cosa que només es pot pressuposar en textos tancats d'aplicació força limitada (recordeu 3.2-3). Sens dubte, és un avantatge computacionalment crucial. La trajectòria que ho

assoleix tanmateix no sols és llarga i complexa per ella mateixa, sinó que transcendeix l'àmbit del lèxic, ja que, en rigor, el diccionari consta de la col·lecció d'entrades de (29) i dels esquemes associats, i aquests valors no sols no es corresponen a res que s'assembli al lèxic morfològicament identificable de cap llengua natural, sinó que no hi són ni en són directament projectables. Ans al contrari, a la vista del que hem comentat a la secció anterior, la dependència conceptual no invalida el processament composicional morfològic i sintàctic, sinó que, almenys en teoria, l'implica. Això comporta afegir un algorisme, generalment força complex, per connectar, avant i enrera, representacions extralingüístiques amb estructures lingüístiques.

5.2 La situació, en aquest punt, es torna paradoxal. Per una banda, l'adopció d'una representació per conceptes transcendeix l'estructura purament gramatical dels textos i assumeix el coneixement dels parlants sobre l'entitat dels referents i les possibles relacions que hi ha entre ells. És una nova dimensió, inèdita a la que hem vist fins aquí, que es troba en el centre mateix de la recerca en IA. Per altra banda, això crea encara nous vincles de relació amb les estructures lingüístiques implicades atès que comporta la comprensió integral dels textos en procés. Els models oracionals no basten i cal recórrer a opcions supraoracionals –en dominis propis de la pragmàtica i la lingüística textual– que puguin interpretar inferències, pressuposicions, implicacions, etc. enmig d'estructures farcides de vaguetats, ambigüitats, elisions i fets indirectes de parla. La paradoxa sorgeix perquè, avui per avui, com més augmenta el poder teòric dels models (en fer-se més complexos i assumir més fenòmens), més en minva el pràctic, en el sentit que el seu abast d'aplicació es redueix quasi proporcionalment.[29]

Una forma habitual no tant de resoldre com de simular aquest repte consisteix a definir un àmbit, uns objectes i una xarxa de possibles relacions entre ells, seguint més o menys el primitiu model de SHRDLU, creat el 1972 per T. Winograd (amb un pàrsing d'ATN). Dintre d'un espai tridimensional determinat s'hi defineixen objectes geomètrics de diversa aparença –com cubs, esferes i piràmides– de dimensions, propietats físiques i situació relativa igualment determinada. El sistema així construït és capaç de reconèixer no sols construccions agramaticals (imprevistes en la producció de regles sintàctiques), sinó també 'impossibles', com una que pretengués col·locar una bola sobre una piràmide.

És, com deia, el domini central de recerca en IA i els seus formalismes peculiars, que comprenen sobretot les xarxes semàntiques amb multitud de modalitats: 'scripts', 'frames', 'mops', 'memettes', etc. Un domini certament apassionant obert a estratègies agosarades (com les xarxes neuronals) que supera de molt l'àmbit de la lexicografia lingüística, tot i que el travessa de cap a cap amb qualsevol dels formalismes que hem esbossat.

(29) Hi ha propostes que no passen de ser el que s'anomena 'gramàtiques de joguina', sense cap altra aplicació pràctica que servir de base d'experimentació o d'exercici pedagògic.

6. SIGLES D'ÚS MÉS GENERALITZAT

IA intel·ligència artificial
LC lingüística computacional
LT lingüística teòrica
PLN processament del llenguatge natural
TA traducció automàtica

RAMON CERDÀ
Universitat de Barcelona

REFERÈNCIES BIBLIOGRÀFIQUES

ALSHAWI, H. (1987) *Memory and Context for Language Interpretation*. Cambridge, Camridge University Press.

BARR, A.; FEIGENBAUM, E. A. (comps.) (1986) *The Handbook of Artificial Intelligence*. Reading Mass., Addison-Wesley. Especialment el Cap. IV, Understanding Natural Language, pàgs 223-321.

BOGURAEV, B. & BRISCOE, T. (1989) *Computational Lexicography for Natural Language Processing*, London, Longman.

CHAFFIN, R. (1988) «The nature of semantic relations: a comparison of two approaches». A: EVENS, M. W. (comp.) (1988) pàgs. 289-334.

DOWTY, D. R. (1979) *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. London, Reidel.

DOWTY, D. R.; KARTTUNEN, L.; ZWICKY, A. M. (comps.) (1985) *Natural Language Parsing. Psychological, Computational, and Theoretical Perspectives*. Cambridge, Cambridge University Press.

EVENS, M. W. (comp.) (1988) *Relational Models of the Lexicon. Representing Knowledge in Semantic Networks*. Cambridge, Cambridge University Press.

GRISHMAN, R. (1986) *Computational Linguistics. An Introduction*. Cambridge, Cambridge University Press.

MARKOWITZ, J. (1988) « An exploration into graded set membership». A: EVENS, M. W. (comp.) (1988) pàgs. 239-60.

MEL'CUK, I. (1988) «The explanatory combinatorial dictionary». A: EVENS, M. W. (comp.) (1988) pàgs. 41-74.

MEYA, M.; HUBER, W. (1986) *Lingüística computacional*. Barcelona, Teide.

NIRENBURG, S. (comp.) (1987) *Machine Translation. Theoretical and Methodological Issues*. Cambridge, Cambridge University Press.

SELLS, P. (1985) *Lectures on Contemporary Syntactic Theories. An Introduction to Government and Binding Theory, Generalized Phrase Structure Grammar, and Lexical-Functional Grammar*. Stanford, Center for the Study of Language and Information. [Traducció en castellà (1989) *Teorías sintácticas actuales*. Barcelona, Teide].

SHIEBER, S. M. (1986) *An Introduction to Unification-Based Approaches to Grammar*. Stanford, Center for the Study of Language and Information. [Traducció en castellà (1989) *Introducción a los formalismos gramaticales de unificación*. Barcelona, Teide].

SLOCUM, J. (1988) *Machine Translation Systems*. Cambridge, Cambridge University Press.

SOWA, J. F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*. Reading Mass., Addison-Wesley.

WINOGRAD, T. (1983) *Language as a Cognitive Process*, Volume I: Syntax. Reading Mass., Addison-Wesley.