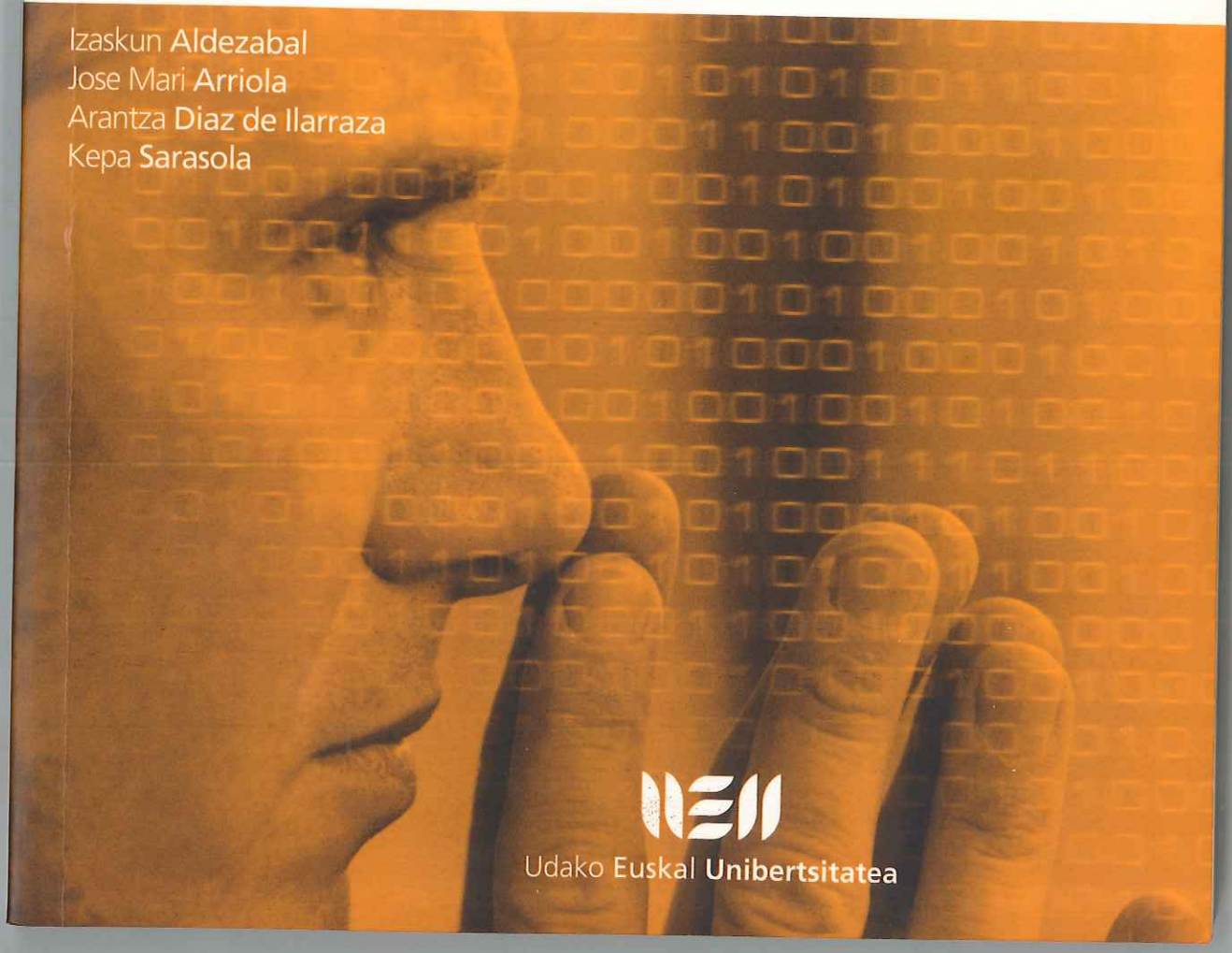
Hztek

graduondokoa

Hizkuntzalaritza Konputazionala

Izaskun Aldezabal
Jose Mari Arriola
Arantza Diaz de Ilarraza
Kepa Sarasola



Udako Euskal Unibertsitatea

Hizkuntzalaritza konputazionala

Izaskun Aldezabal

Jose Mari Arriola

Arantza Diaz de Ilarraza

Kepa Sarasola

Udako Euskal Unibertsitatea

Bilbo, 2005

ISBN: 84-8438-065-3
Lege-gordailua: BI-1407-05

Inprimategia: RGM, Bilbo
Azalaren diseinua: Iñigo Ordozgoiti
Hizkuntza-zuzenketen arduraduna: Ander Altuna Gabiola

Banatzaileak: UEU. Erribera 14, 1. D BILBO telf. 946790546 Faxa. 944793039
Helbide elektronikoa: argitalpenak@ueu.org
www.ueu.org

Zabaltzen: Igerabide, 88 DONOSTIA telf. 943310301

Hitzaurrea

Hizkuntza-Teknologiak funtsezkoak dira Informazio eta Komunikazioaren Gizartea esaten dugun horretan. Epe ertainean pertsonen eta makinen arteko komunikazioa, hainbat aplikazio informatikotan, geure hizkuntzan egin ahal izango dugu, ez makinaren hizkuntzan. Tresna mugatuak izango dira, eta beti errore maila batekin, baina, hala ere, laguntza ederra emango digute. Gaur egun eskura ditugu horrelako zenbait hizkuntza-aplikazio, halanola: ortografia-zuzentzaileak eta estilo-zuzentzaileak, hiztegi-kontsultak on-line, itzulpen-laguntzak, Interneterako bilatzaileak, hizketa testu bihurtzen duten sistemak, testua irakurtzen dutenak edo bigarren hizkuntza ikasteko sistemak.

Baina horrelako sistema gehienek ingeleserako balio dute, ez beste hizkuntzetarako. Beste hizkuntzetako hiztunok ahalegin handia egin behar dugu atzean ez gelditzeko, are gehiago euskara bezalako hizkuntza txikikoek.

Aztertzen badugu Euskararen Softwarearen Katalogoa (http://www.ueu.org/softkat), hizkuntzaren prozesamenduari lotuta dauden 36 aplikazio aurkituko ditugu. Ez da hutsaren hurrengoa, baina bai oso gutxi gaztelaniarako, frantseserako edo, batez ere, ingeleserako eskuragarri dauden ehunka programa eta baliabiderekin alderatzen badugu.

Euskararako baliabide linguistikoak sortu beharko dira: testu-bilketa (corpus orokorra, berezituak eta eleanizdunak, ahal bada 100 milioi hitzekoak). Hiztegi elektronikoak eta hitzen sailkapen orokorra adieraren arabera. Aplikazio berriak plazaratu beharko dira: informazio-bilatzaileak (eguraldi-parteak, burtsa, kirolak, berriak, bideo-eskaerak, irudi-eskaerak...), domotika, itzulpen-laguntzak, irakaskuntza-sistemak (e-learning), elbarrientzako laguntzak, telebista digitala, elkarrizketa-sistemak, multimedia-sistemak...

Erakunde ofizialetatik bultzatu egiten da ikerketa-lerro hau. Europar Batasunak I+Grako aurrekontuaren % 3,77 (564 milioi euro) bideratu du "Multimedia edukiak eta Tresnak eta Hizkuntza Teknologiak" alorrerako. Eusko Jaurlaritzak 2001. urtetik ingeniaritza linguistikoa ikerkuntzarako lerro estrategikoen artean hartu du Zientzia, teknologia eta berrikuntzarako planetan.

Hizkuntzaren teknologian euskarak, bere kabuz, aurrera egitea nahi badugu, sortzaile euskaldunak behar ditugu. Euskarak eta euskaldunok, Informazio eta Komunikazioaren Gizartearen erronkari erantzun ahal izateko, profesional euskaldunak behar ditugu. Formazio-plan horretan kokatzen da UEUren papera. Udako Euskal Unibertsitateak heziketa-behar hori bete ahal izateko, hainbat unibertsitate eta enpresatako adituak bilduz, 2001/2002 ikasturtean HIZKUNTZA-TEKNOLOGIAK masterra jarri zuen martxan. Ikasketa berri honek UPV/EHUren oniritzia jaso zuen, eta beraz, unibertsitateak titulua bere babesarekin zabaldu du. Udako Euskal Unibertsitatearen eta EHUko Lengoaia eta Sistema Informatikoak sailaren arteko lankidetzak fruitu onak eman ditu, aurten Euskal Filologia saila ere sartu da.

Nazioartean puntako mailan mugituko den industria sendoa sor dezakegu. Ikertalde, enpresa eta erakunde ofizialetatik ahaleginak egiten ari gara asmo horiek gauzatzeko. Aldi berean Hizkuntzaren Teknologiarako adituak prestatu behar dira, horixe lortu nahi du UEUren HIZTEK masterrak. Informatikariak, ingeniariak, hizkuntzalari eta filologoak dira harrobi horretarako lehengaia.

Testuinguru horretan, adituak prestatzeko helburu horrekin, liburu hau UEUk hizkuntza-teknologien inguruan argitaratzen duen bigarrena da. Lehenengo liburuan morfologia landu genuen, atal zehatz bat. Bigarren honetan, ordea, hizkuntza-teknologien ikuspuntu orokorra aurkezten dugu bi atal nagusitan: ereduak eta produktuak.

Hizkuntza ulertzeko, aztertzeko eta azaltzeko saioan erabiltzen diren ereduak hartzen ditu liburuko lehen atal nagusiak, lau kapitulutan banatuta: morfologia, sintaxia, semantika eta pragmatika.

Bigarren atalean produktuak azaltzen dira: 6. kapituluan erabiltzaile arruntarentzako aplikazioak (zuzentzaileak, hiztegi-kontsultak on-line, itzulpen automatikoa, testu-masa handiak tratatzekoak, hizketaren tratamendua eta interfazeak); 7. kapituluan hizkuntza-softwarea sortzen dutenentzako tresnak (analizatzaile morfologikoak edo sintaktikoak adibidez); eta 8. kapituluan edozein aplikazio edo tresna garatzeko behar-beharrezkoak diren hizkuntza-baliabide eta oinarri linguistikoak (corpus, hiztegiak eta ontologiak adibidez).

Donostian, 2005eko maiatzaren 20an

Aurkibidea

Lehenengo atala

ATARIKOAK

1 Sarrera

Liburu honek hizkuntza-teknologiako gaien sarrera izatea du helburu, alorrari ikuspegi orokorra emanez.

Motibazio nagusia hau da: alor honek bere baitan dituen bi diziplinen —hizkuntzalaritzaren eta teknologiaren— arteko loturen berri ematea eta lotura horiek egiteak dakarren konplexutasuna agerian jartzea. Hori dela-eta, bi diziplinak uztartzeko bi alorretako ikertzaileen elkarlana ezinbestekoa dela erakutsiko dugu. Bakoitzaren materialak, jakinduria eta eskarmentua elkartuz, diziplina arteko arlo honek hizkuntzaren azterketa-bideetan aurrera egingo du.

Liburu honen izenburuan *Hizkuntzalaritza Konputazionala* terminoa erabili da, egileon ustez adar hauxe baita Hizkuntzari eta Teknologiari, biei, garrantzia emanez, uztarketa-lana egokien bideratzen duena. Ordenagailuak hizkuntza uler dezan beharrezkoa duen ezagutza formalizatzea da bilatzen dena. Liburu honetan, hortaz, ezagutza mota horiek zeintzuk diren eta nola formalizatu diren —edo behar diren— aztertzen da.

Hizkuntza-ezagutza ikuspegi konputazionaletik formalizatzea, ordea, ez da zeregin erraza. Bost urteko edozein ume hitz egiten eta ulertzen ondo moldatzen denez, hizkuntza erabiltzea lan erraza dela pentsatzen dugu, baina hori ez da horrela. Hizkuntza sortzea eta ulertzea oso prozesu konplexuak dira, eta gaur egungo ordenagailuak urrun ikusten ditu giza adimenaren ahalmen linguistiko orokorrak. Baina horrek ez du esan nahi hizkuntza lantzeko tresna automatikoak utopia direnik, hizkuntzaren oinarrizko ezagutza minimo batekin laguntza interesgarriak eskain daitezkeelako. Helburu horretan, testuaren zailtasun mailaren arabera emaitzak asko alda daitezke. Horrela, emaitza probetxugarriak lortzeko, ordenagailuaren lana aztergai espezifiko eta mugatu batean kokatu behar da. Etorkizunean, hala ere, aplikazio mugatuko sistemak bilduz, lor litezke ahalmen handiagoko sistema berriak, baina egun ibili dabiltzan aplikazioek helburu espezifikoak dituzte.

Hizkuntza ulertzeko, aztertzeko eta azaltzeko saioan, hizkuntzalaritza orokorrean eta konputazionalean erabiltzen diren ereduak ez dira berdinak izan, ezin berdinak izan ezinbestean. Konputagailuak hizkuntza prozesatze aldera, konputagailuari lagungarri gerta ahal zaizkion eredu aplikatuak egiten dira batik bat. Horrek hizkuntzaren formalizazio zorrotza eskatzen du. Izan ere, ikerketak erakutsi du hizkuntzalaritza orokorrean erabiltzen diren ereduak, bere horretan ezin egokitu zaizkiola zehatz-mehatz ordenagailuari. Horregatik, hizkuntzalaritza konputazionalak bere ereduak sortu edota egokitu behar izan ditu, ordenagailuarentzat egokiagoak izateko. Horregatik, hizkuntzalaritza konputazionalak ere, konputazionalki tratatu ahal izango diren eredu teoriko batzuk aukeratzen ditu, eta hortik aplikazioak ateratzen ditu. Ondorioz, hizkuntzalaritza konputazionalak bi alderdi ditu: teorikoa —hizkuntza tratatzeko ereduak— eta praktikoa —aplikazioak, helburu konkretuei begirakoa—.

Bestalde, alde aplikatutik zein teorikotik hizkuntzaren azterketari ekiterakoan, anbiguotasuna da arazo nagusietako bat.

Anbiguotasuna hizkuntzaren maila guztietan badago ere, hizkuntzalaritza orokorrean eta konputazionalean anbiguotasunaren kontzeptua ez da berdin erabili. Horrela, ikuspegi konputazionalean anbiguotasun lexikoa ebaztea oso garrantzizkoa da (morfologikoa lehenengotik eta semantikoa gero); aldiz, teorikoarentzat ez da horren premiazkoa. Zergatik da horren garrantzitsua konputazionalarentzat maila hori? Lehenik eta behin, hitz batek bere testuinguru guztietan izan ditzakeen interpretazioak zehaztu behar dira. Esaterako, *zituen* hitza aditz laguntzailea da *Liburuak ekarri zituen* moduko testuinguru batean, baina baita *zitu* izenaren ('fruitu, emaitza') genitibo plurala ere, *Zituen artean bada desberdintasunik* moduko testuinguruetan. Ondorioz, dagokion interpretazioa aukeratzeko, lehenik eta behin intrepretazio horrek egon egon behar du. Behin interpretazio guztiak edukita, testuinguruari dagokiona aukeratu behar da. Hots, hitz hori desanbiguatu beharra dago, hurrengo analisi sintaktikoak taxuz egin nahi baditugu. Azken finean, aurrena analizatzaile morfologikoari hitz bat analizatzeko beharrezko informazioa ematen zaio, eta ondoren, horixe bera bueltan etortzen da, dagokiona aukeratzeko. Horregatik, askotan, hitz baten analisi morfologikoan imajinatu ezinezko sorkuntzak proposatzen dira (lehen kolpean, *zitu + en* bera izan daiteke horrelako bat), baina testuinguru jakinen batean behintzat zilegi dela pentsatu behar da. Eta hori ere hizkuntza lantzea da; hizkuntzak eskaintzen dituen aukeren jabe egitea. Ideia hau giltzarria da hizkuntzalaritza konputazionalean.

Hala, anbiguotasun morfologikoaren (lexikala, hitzarena) azterketa eta desanbiguazioa, lan morfologikoen azken urratsa da eta era berean sintaktikoen lehenengoa. Analisi sintaktikoari ekiterakoan hori da konponduta egon behar duen lehenengo gauza.

Analisi sintaktikoak edo estrukturalak, semantikoak eta pragmatikoak izaten dira hizkuntzalari teorikoen buruhausteak, eta esango genuke maila sintaktikoan hasten dela anbiguotasuna haientzat. Konputazionalean ere, noski, eta anbiguotasun mota guztiak korapilatsuak badira konpontzeko, zailtasun maila *in crescendo* doa mailaz maila, baina lexikotik eta morfologikotik hasita.

Zer gertatzen da euskaraz eta euskararentzat egiten den hizkuntzalaritza konputazionalarekin? Faktore asko dago euskararen azterketa konputazionala bereizi egiten dutenak inguruko hizkuntzen azterketatik.

- Alde batetik, euskara hizkuntza minorizatua izanik, azterketotarako behar izaten diren oinarrien falta sumatzen da, corpusena, esate baterako. Azken finean, merkatu-interes kontuengatik garatu dira horrenbeste ingelesarentzako errekurtsoak, eta minorizatua izateagatik edo ezagutza ofizial ezarengatik merkatu-interes falta nabaritu dute euskara bezalako hizkuntzek.
- Bestetik, estandarizazio berantiarrak ere normalizazio-prozesua motelagoa izatea ekarri du, eta horrek ikerketa guztiak beste hizkuntzetan baino beranduago gertatzea ekarri du. Ingelesaren gainean 60-70eko hamarkadetan hasi ziren inguruko lanak, 80ko hamarkadan *parsing*-ari buruzkoak. Espainolari buruzko lanak 80koan eta euskarazkoak 80koaren bukaeran-90ekoan.
- Bi puntu hauei hizkuntza-tipoarena gehitu behar zaie. Horrela, euskara, familia indoeuroparretik kanpo kokatzen denez, ezaugarri asko eta asko ez ditu inguruko hizkuntzekin konpartitzen. Horrek esan nahi du hizkuntza horietarako erabiltzen diren hainbat eredu formal eta irtenbide ezin zaizkiola zuzenean euskarari aplikatu.

Hala ere, azken 10-15 urteotako lanarekin alor honetan iritsi dugun maila altua da dudarik gabe. “Berandu baina seguru” horixe esan daiteke: 10-15 urteotan izugarrizko lana egin da eta beste hizkuntzetan egiten diren lanokin konparatuta, batere inbidiarik gabe ibiltzeko moduan gaude... baina oraindik ingelesarekiko alde handia dago, batez ere baliabideen inguruan.

Motibazioa aztertuta, liburuaren helburuak, laburki, honela zehatz ditzakegu:

- Hizkuntzalaritza konputazionalaren nondik norakoen berri ematea eta zer den ulertaraztea.
- Alor honen barruan sortu diren zenbait oinarrizko kontzeptu eta behar diren ezagutza motak azaltzea.
- Alor honetatik sortzen diren produktuen berri ematea.
- Ikuspegi orokorra zein euskarari loturikoa ematea.

Azpimarratu behar da, hemen ematen dugun informazioa lan-alorraren ikuspegi orokorra izateko baliagarria den arren, liburua argitaratu eta zenbait urtetara zaharkitua geratuko dela hein batean, batez ere, ematen diren web guneei dagokienez, eta azaldu ditugun oinarri eta tresnen garapenari dagokionez. Nolanahi ere, aspalditik zegoen hutsunea betetzera datorrela uste dugu.

1.1 Oinarrizko kontzeptuak

Esan dugun bezala, hizkuntza-teknologietan bi motibazio nagusi biltzen dira: teknologikoa eta linguistikoa. Motibazio teknologikoari dagokionez, helburua konputazio-sistema adimentsuak garatzea da, hala nola, datu-baseei hizkuntza naturalez galdetzeko interfazeak, itzulpen automatikorako sistemak, testuen analisirako tresnak, hizketaren tratamendurakoak, etab. Motibazio linguistikoari dagokionez, berriz, alderdi linguistikoari erreparatzen zaio, eta bereziki, hizkuntzaren modelizazioa lantzen da. *Hizkuntza-teknologiak* termino honen baitan, beraz, hizkuntzalaritza teorikoa eta praktikoa, informatika eta adimen artifiziala biltzen direla esan genezake, edo, bestela esanda, alde teorikoa eta ingeniaritzari dagokiona ere bai.

Ondoren, diziplina bera izendatzeko erabiltzen diren terminoak zehaztuko ditugu.

1.1.1 IL, HK eta LNP

Alderdi teorikoa eta teknologikoa kontuan harturik, hizkuntzaren teknologiaz hitz egiten denean, honako termino hauek azaltzen zaizkigu: **ingeniaritza linguistikoa**, **hizkuntzalaritza konputazionala** eta **lengoaia naturalaren prozesamendua**. Esan dezagun, hala ere, gure asmoa hurbilpen horiek bereiztea bada ere, askotan, arlo hauetan ibiltzen direnek ere zenbait testuingurutan bata zein bestea sinonimo gisa erabiltzen dituztela. Kontzeptu horien arteko mugak lausoak dira. Beraz guk emango ditugun ñabardurak modu malguan hartu behar dira.

Bestalde, termino horien baitan, lantzen denari erreparatuz gero, denek dute helburu nagusi bera: ordenagailuaz baliatuz hizkuntza tratatzea. Kontua da, hizkuntza tratatzerakoan nagusitzen den ikusmoldearen arabera desberdintzen direla:

- **Ingeniaritza linguistikoa** (IL)/(*Language Engineering,* LE). Hizkuntzari buruzko ezagutza batez ere aplikazioetara eta produktu komertzialetara zuzenduta dago. Hizkuntza ezagutzeko, ulertzeko,

interpretatzeko eta sortzeko gai diren sistema informatikoak garatzea du jomuga. Honi hizkuntza-teknologia ere (*Human Language Technology,* HLT) esaten zaio.

- **Hizkuntzalaritza konputazionala** (HK)/(*Computational Linguistics,* CL). Ikuspegi abstraktuago batetik ekiten dio hizkuntzaren modelizazioari ordenagailuek hizkuntza uler dezaten. Hau da, hizkuntza formalizatzen dute ordenagailuek ulertu ahal izateko moduan.
- **Lengoaia Naturalaren Prozesamendua** (LNP)/(*Natural Language Processing,* NLP). Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloari Lengoaia Naturalaren Prozesamendua (LNP) esaten zaio, eta, batez ere, erabiliko diren teknika informatikoei erreparatzen die: ezagutza linguistikoa nola adierazi konputagailuan, nola erabili ezagutza hori (algoritmoak, estrategiak, inferentziak sortzeko metodoak, etab.), nola uztartu programetan ezagutza linguistikoa eta hizketa-gaiari dagokion ezagutza, nola banatu tratamendu linguistiko osoa modulu sinpleago eta independentetan horietako bakoitza egingarriagoa izan dadin...

Bereizi behar izatekotan zer bereizi beharko litzatekeen irudikatu dugu, eta ez hitz horien aipamen guztiak modu ziurrean bereizteko araua.

Beste alde batetik, hizkuntzaren tratamendu automatikoaren barruan badira beste bi kontzeptu, bereizi izan direnak: hizkuntza idatziaren eta hizkuntza mintzatuaren tratamenduak. Azken horren azterketa konputazionalari erreferentzia egiteko *hizketaren tratamendua* terminoa erabili ohi da. Horregatik, *hizkuntzaren tratamendua* terminoa, arrunki, hizkuntza idatzizkoarekin lotu ohi da. Bereizketa hori gertatu izan da urtetan arlo diferenteak izan direlako, problematika eta metodologia oso bestelakoak erabili dituztelako. Baina mende bukaeratik aurrera atal bi horiek gero eta modu koordinatuagoan lantzen ari dira. Egun, askoz harreman handiagoa dago hizketa eta testua aztertzen dituzten ikertzaileen artean, lehen hain bereizita lan egiten zuten ikerketa-komunitateek elkarren beharra nabaritu baitute.

1.1.2 Arlo konputazionala hizkuntzalaritzan

Hizkuntzalaritza orokorrak teoria gramatikal dotorea, murriztua eta unibertsal linguistikoen berri emango duena du helburu; hizkuntzalaritza konputazionalak, berriz, aplikagarritasuna duen sistema eraiki nahi du, hots, egitura linguistikoa eraginkortasun konputazionalarekin prozesatzea du jomuga.

Hizkuntzalaritza orokorra hiztunen konpetentzia aztertzeaz arduratzen da bereziki, eta datuak erdiesteko, batez ere, introspekzioa du iturri nagusi. Ondorioak dedukzio-metodoen bidez erdiesten ditu. Hizkuntzalaritza konputazionala, berriz, erabilera linguistikoan zentratzen da, komunikazio-egoera errealetatik datuak erdietsiz. Ikertzeko, dedukzio- zein indukzio-metodoak baliatzen ditu. Oro har, esan dezakegu hizkuntzalaritza orokorra teoriaren dotoretasunaz arduratzen dela gehiago, eta hizkuntzalaritza konputazionala, sistemaren erabilgarritasunaz eta eraginkortasunaz.

Dena dela, hizkuntzalaritzaren baitan, teoriak garatzeko datu objektiboetatik abiatzen diren hurbilpenek gero eta indar handiagoa dute. Ildo honetatik, aurrerapen handiak lortu dira ordenagailuen erabilerari esker, esate baterako, corpusean oinarritutako hizkuntzalaritzan.

Bestalde, Sparck Jones-en lanean (1996:14) esaten denez, oro har, hizkuntzalaritzaren, eta bereziki hizkuntzalaritza teorikoaren eragina HKn oso ahula izan da. Horrez gain, informazio-teknologiaren eragina hizkuntzalaritzan, HKtik kanpo oso zaila dela aurkitzen ere esaten du, eta uste duela hizkuntzalaritzak asko duela irabazteko arlo konputazionaletik:

> *"..., there is much for linguistics to gain from looking both at how computation does things and at what it finds".*

Ikusmoldearen gorabeherak alde batera utzita, denen oinarrian ikerkuntza dago, **hizkuntza naturalaren tratamendu automatikoa**ren arloko ikerkuntza, alegia. Guk liburuan aztertuko dugunari *hizkuntzalaritza konputazionala* deituko diogu oro har, eta horren bidez alderdi teorikoak eta teknologikoak, biek, garrantzia dutela azpimarratu nahi dugu.

Ikuspuntu horretatik hizkuntzalaritza konputazionalaren xedea guretzat honako hau izango da: hizkuntza-ulermenaren eta sorkuntzaren teoria konputazional ulergarri, taxutu eta linguistikoki motibatua eraikitzea.

1.1.3 Datu linguistikoak vs programak

Hizkuntzaren tratamendu automatikoaren arloan, edozein dela ere aplikazio mota, komeni da bereiztea aplikazio ororen oinarrian dauden bi osagai :

- **Datu linguistikoak.** Baliabide lexikal hauetan ager daitezte: lexikoiak, hiztegiak eta gramatikak, besteak beste.
- **Programak.** Datu linguistiko horien (hots, baliabide lexikaletan ditugunak) gainean aplikatzen diren programak informazio linguistikoa prozesatu ahal izateko: analizatzaileak, desanbiguatzaileak, programa estatistikoak...

Horiek dira aplikazioen oinarriak; hasieratik argi utzi behar dugu, ordea, ez garela ariko programei[1] buruz.

Hizkuntzaren tratamendu automatikoaren lehenengo urteetan, programak eta ezagutza linguistikoa (hots, datu linguistikoak) ez ziren bereizten. 80ko hamarkadaren erdialdera, bi informazio mota horien bereizketa hasiko da. Horrela, ezagutza linguistikoaren (gramatika batean eta lexikoan kodetuta) eta datu horien gainean lan egingo duten prozedurak bereiziko dira. Hartara, bereizketa horren abantailak hauek izango dira:

- **Ekonomia:** deskribapen linguistiko berak balia daitezke prozesamendu desberdinetarako (analisirako zein sorkuntzarako, esaterako).
- **Baliotasun teorikoa:** deskribapen gramatikala programaren exekuzio-prozeduretatik independentea baldin bada, hizkuntzalaritza teorikoaren emaitzak hobeto balia daitezke.

---

[1] Informazio linguistikoa prozesatu edo tratatu ahal izateko modelo eta algoritmoei buruzko oinarrizko informazioa jaso nahi duenak jo beza liburu honetara: Daniel Jurafsky & James H. Martin, *Speech and Language Processing,* 5-6. or.

- **Deklaratibotasuna:** ezagutza linguistikoa modu deklaratibo batez deskriba daiteke, hau da, kasuan kasuko hizkuntzari dagozkion egitura linguistikoak adieraziz.

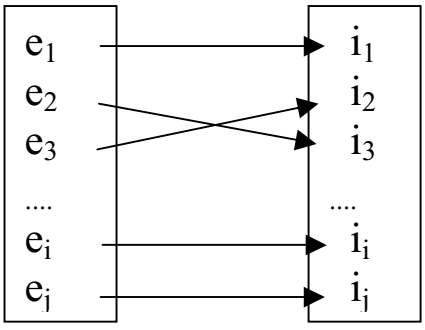
1.2 Hizkuntza prozesatzeko arazoak

Hizkuntza naturalaren tratamendu konputazional osoa ezinezkoa da, etengabeko aldaketak eta erabilera mugaezinak ditu eta. Atal honetan konplexutasun hori sortzen duten arazo nagusiak azaltzen saiatuko gara: anbiguotasuna, errepresentazio konplexuaren beharra eta tratamendu independentea onartzen ez duten ezagutza mota asko denak batera aldi berean erabili beharra.

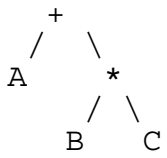
1.2.1 Anbiguotasuna

Hizkuntzaren berezko ezaugarrien artetik, tratamendu automatikoari begira, anbiguotasuna da arazo gehien sortzen dituena. Hitz batek adiera bat baino gehiago izan ditzake, batzuk gainera kategoria morfologiko desberdinetakoak dira; esaldi batek analisi sintaktiko bat baino gehiago onar ditzake; egoeraren arabera esaldi berak esangura desberdinak eskain ditzake. Guk, gizakiok, gehienetan, egoeraren arabera ondo dakigu une bakoitzean adiera, kategoria eta analisi egokiak aukeratzen, askotan gainera hainbeste aukera burutik ere ez zaizkigu pasatzen, baina hori ez da lan samurra konputagailuarentzat. Izan ere, ordenagailuan pilatutako informazio guztia (lexikala, morfosintaktikoa, semantikoa nahiz pragmatikoa) prozesatzean, askotan pentsatu ezinezko emaitzak sortuko dira (sarreran aipatu dugun adibidea ekarriz, nork pentsatuko luke lehenengotik, aditzaz gain *zituen* forma *zitu* izenaren genitibo plurala dela?).

Tratamendu automatikoari begira errazena litzateke esaldiaren eta bere interpretazioaren arteko korrespondentzia bijektiboa (bat-bat) izatea, informatikako lengoaietan gertatzen den bezala. Horrela, esaldi bakoitzarentzat interpretazio posible bakarra egongo litzateke, eta interpretazio bat adierazteko esaldi bakarra sor liteke.



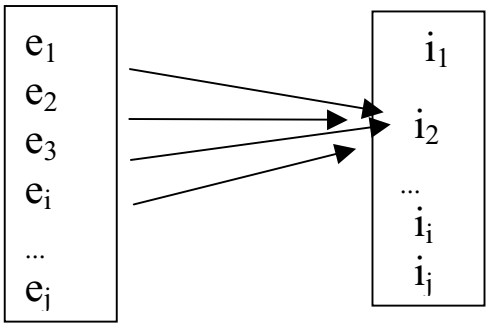
Adibidez, programazio-lengoaia baten bidez adierazpen matematikoa idatzi nahi dugunean korrespondentzia bijektibo bat ezar daiteke adierazpen matematikoaren eta bere zuhaitz-errepresentazioen artean. Esaterako `A+B*C` adierazpen matematikoari honako errepresentazio hau dagokio;



Baina hizkuntzaren erabilera zailagoa da. Eta zailagoa da alde bietatik: batetik, hainbat esaldik interpretazio bera izan dezaketelako (*hainbat-bat* erako korrespondentzia), eta, bestetik, esaldi berak interpretazio bat baino gehiago izan ditzakeelako (*bat-hainbat* erako korrespondentzia).

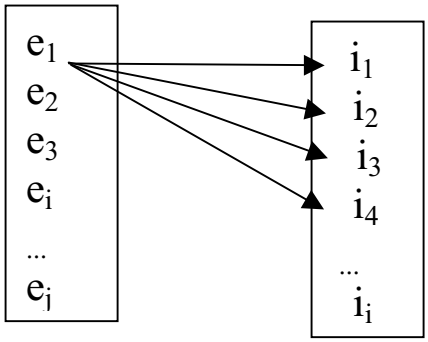
Batetik, *hainbat-bat* erako korrespondentzia gertatzen da esaldiak ulertzerakoan, hainbat esaldik interpretazio bera dutenean; adibidez, interpretazio bera lortu beharko litzateke unibertsitateko araudiak biltzen dituen datu-base bati honako galdera hauetako bat, edozein, eginda:

- *Klaustroko hauteskundeei buruz dagoen guztia nahi dut.*
- *Ezer daukazu klaustroko hauteskundeei buruz?*
- *Emango didazu klaustroko hauteskundeen arautegia?*
- *Mesedez, klaustroko hauteskundeena!*



Eta bestetik, *bat-hainbat* erako korrespondentzia ere gerta daiteke esaldiak ulertzerakoan, esaldi bakar batek testuinguru desberdinetan hainbat interpretazio dituenean; adibidez, ondoko bi esaldiek interpretazio bat baino gehiago eduki dezakete:

- *Bosgarren udako unibertsitatea*
  1. *(Bosgarren udako) unibertsitatea*. Bosgarren udan antolatu zen unibertsitatea
  2. *Bosgarren (udako unibertsitatea)*. Bosgarren aldian antolatu zen udako unibertsitatea
- *I see a man in the park with the telescope*
  1. Interpretazio ohikoena: Gizon bat ikusten dut parkean teleskopio batekin.
     I see (a man)
     (in the park)
     (with the telescope)
  2. Teleskopio batekin parketik dabilen gizon bat ikusten dut.
     I see (a man (in the park))
     (with the telescope)
  3. Teleskopio bat duen parkean gizon bat ikusten dut.
     I see (a man)
     (in the park (with the telescope))
  4. Teleskopio bat duen parketik dabilen gizon bat ikusten dut.
     I see (a man (in the park (with the telescope))



Zailtasun horiexek bihurtzen dute anbiguotasuna LNPko arazo nagusietako bat. Hala ere, esan dugun bezala, ez da soilik LNPren arazoa, izatez hizkuntza baita anbiguoa. Horren ildotik, ikusi besterik ez dago zenbait hizkuntzalariren hitzak gai honen inguruan:

> *"Que la ambigüedad es connatural al lenguaje común –a lo que llamamos lengua a secas—en cualquiera de sus variadísimas especies es un hecho tan conocido que no hace falta apelar a refinadas técnicas dialécticas y retóricas para traer a los incrédulos al buen camino (...). La ambigüedad es, sin lugar a dudas, uno de los universales más patentes del lenguaje natural (...)"* (Michelena, 1972).
>
> *"Por lo que al lenguaje atañe, el sistema (en sus diversos niveles) tiende a ser distintivo, incluso con notables grados de redundancia. A pesar de ello, queda abierta la puerta para que se introduzca la posibilidad de ciertos sincretismos (morfológicos, léxicos y sintácticos) con los que se crea una perturbación en el proceso comunicativo".* (Tusón, 1975:325).

Gai honi buruz asko idatzi da, dudarik gabe, eta ikuspuntu diferenteetatik planteatu da arazoa. Izan ere, oso eremu zabala denez, mota askotako alterazio linguistikoak egon daitezke definizio horren barruan. Ikus bestela Padró-ren tesian (Padró, 1997):

> *"Ambiguity in natural language is mainfold. We find part-of-speech ambiguity (e.g. past vs. Participle in regular verbs), semantic ambiguity in polysemic words, syntactic ambiguity in parsing (e.g. PP-attachment), reference ambiguity in anaphora resolution, etc.".*

Horrela, alderdi teorikotik zein konputazionaletik azter daiteke anbiguotasunaren arazoa. Anbiguotasun morfologikoa, sintaktikoa, semantikoa eta pragmatikoa ere trata daitezke.

Karlsson-ek (Karlsson *et al.,* 1995) proposatzen duen eskemari jarraituz, hiru multzo nagusitan bereizten dira azaleko mailako anbiguotasun motak: gramatikala (egiturazkoa edo sintaktikoa ere deitua), semantikoa eta pragmatikoa.

Anbiguotasun semantikoak direla eta, ezagunena eta gehien gertatzen dena polisemia dugu edo bestela deituta "anbiguotasun lexikala" (Cristal, 1991, s.v. *ambiguity*). Oso aztertua izan da fenomeno hori hurbilpen konputazionaletan, lexikografia konputazionalaren eremuan batez ere; ingeleserako, erreferentzia hauek ditugu, besteak beste: Kelly-Stone, 1975; Hirtst, 1987; Cottrell, 1989; Ravin, 1990; Rigau, 1999. Euskarazko lexikografia konputazionalean ere anbiguotasun lexikalak bere tokia izan du; Agirre (1999) dugu erreferentzietako bat. Polisemiaz gain, badira anbiguotasun semantikoan kasu gehiago, baina hemen ez dugu gehiago sakonduko berorietan.

Era berean, anbiguotasun pragmatikoa aipatu baino ez dugu egingo. Anbiguotasun pragmatikoa ebazteko denbora- eta espazio-testuinguruaren ezagutza behar da, edota informazio metatestuala. Ingelesez, horrelakoak tratatu izan dituztenen artean hauek ditugu: Litman-Hirchberg (1990) eta Hinkelman-Allen (1989).

1.2.2 Errepresentazioaren konplexutasuna

Hizkuntza-teknologian, lortu nahi den aplikazioa edo tratatu behar den mezua sinplea bada, konputagailu barruan esaldiak errepresentatzeko behar den adierazpidea sinplea izan daiteke. Baina gizakion ulermena simulatu nahi bada, orduan oso errepresentazio konplexuak behar dira. *Adimen artifiziala* deritzon arloan hainbat proposamen sortu dira ezagutza asko bildu eta erabili ahal izateko, baina oraindik utopia hutsa da pentsatzea konputagailu batek erabiliko duela pertsona batek eguneroko bizimoduan behar duena. Beraz, eraiki nahi den aplikazioak hizkuntzaren erabilera sinplea egiten badu, posible izango da martxan jartzea, baina konplexua bada (ezagutza asko, planifikazioa, dedukzioa, ohiko egoeren tratamendua...) oso zaila izango da hizkuntza erabiltzea gizakiok egiten dugun modura.

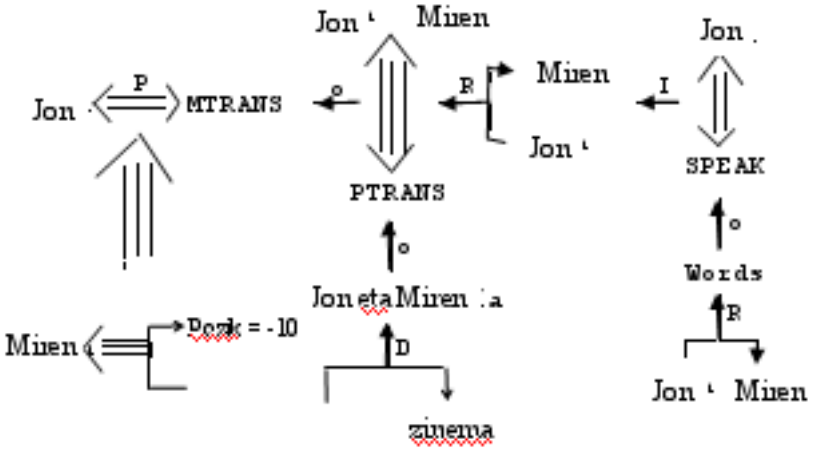
Aplikazio sinplearen adibidea. Arestian aipatu dugun unibertsitateko araudiak biltzen dituen datu-base bati galderak egitea hitz gako batzuk detektatuz egin daiteke, eta gero hitz horiek datuetan bilatuz. Horretarako aski izan daiteke hitz gakoen errepresentazioa, besterik ez da behar.

Esaldia:
Ezer daukazu klaustroko hauteskundeei buruz?
Lortu behar dena:
(SEARCH KEYWORDS= HAUTESKUNDE & KLAUSTRO)

Errepresentazio konplexuaren adibidea. Testu bat emanda, edukiari buruzko galderei erantzuteko gai diren eta era berean dedukzioak egiteko ahalmena exijitzen duten galderei erantzuteko gauza den sistema. Adibidez:

Testua:
Jonek Mireni esan zion zinemara berarekin joango zela
Miren poztu egin zen
Galdera:
Zergatik poztu zen Miren?

Horrelako sistema batean esaldien adierazpide oso konplexua behar dugu. Adibidez, Schank ikertzailearen ereduak esaldi bakoitza diagrama baten bidez adierazten du. Esaldiko aditza ekintza baten bidez adierazten da eta bere osagarriak ere (objektua, subjektua, nondik, nora...) grafikoki, erlazioen bitartez azaltzen dira. Gainera 12 ekintza posible baino ez dago (mugimendu fisikoa, mugimendu psikologikoa...), eta beraz, aditz bakoitzari ekintza primitibo horietako bat dagokio proposatutako ereduaren arabera. Honela errepresentatu beharko lirateke goiko esaldiak eta haien arteko erlazioa.



1. irudia. Esaldietan aurki ditzakegun harremanen errepresentazioa

Eskema horretan adierazi da Jonek hitz egin zuela (SPEAK ekintza). Ekintza horren objektua ("o" erlazioa) hitzak zirela. Hitz horiek Mirenengandik Jonengana joan zirela ("R" erlazioa). Jonek esan zuen mugimendu fisiko bat gertatu zela (PTRANS ekintza, Physical TRANSlation). Mugimendu fisikoa izan zen Jonek eta Mirenek beren burua ("o" erlazioa) zinemara ("D" erlazioa) eramatea. Eta errepresentatu behar izan da bigarren esaldia (*Miren poztu egin zen)* lehenengoaren ondorio gisa gertatu zela: alegia, Jonek aldaketa psikologiko bat (MTRANS ekintza bat, Mental TRANSlation) sortu zuela Mirenengan, horren poztasun-egoera mailarik altuenera igo zela (-10). Eta hori ez dago esplizitu testuan, deduzitu egin behar izan da.

Beraz, gizakiontzat ohiko diren hizkuntza-tratamendu batzuk ordenagailu bidez egin nahi ditugunean oso errepresentazio konplexuak erabili behar dira, eta oraindik konputagailuek ez dute behar besteko tresna sendorik horrelako zailtasunekin egoki jokatzeko.

1.2.3 Tratamendu independentea onartzen ez duten ezagutza mota asko erabili behar dira

Arazo handiei ekin nahi diegunean, *banatze-estrategia* aplikatzea egokia izaten da. Estrategia horren arabera, problema handia hainbat azpiproblema independentetan banatzen dugu, ebazteko errazagoak direlakoan. Azpiproblemetarako lortuko ditugun ebazpenak, geroago, era egokian konbinatuko beharko ditugu problema osoaren ebazpena lortzeko. Estrategia hori oso lagungarria izaten da, baina aplikagaitza izaten da hizkuntzaren tratamendu automatikoan, erabili behar diren ezagutza motek elkarren artean mendekotasun handia izaten dutelako, hau da, azpiproblema independenteak bereiztea oso zaila delako. Saiatuko gara hori erakusten adibide baten bitartez:

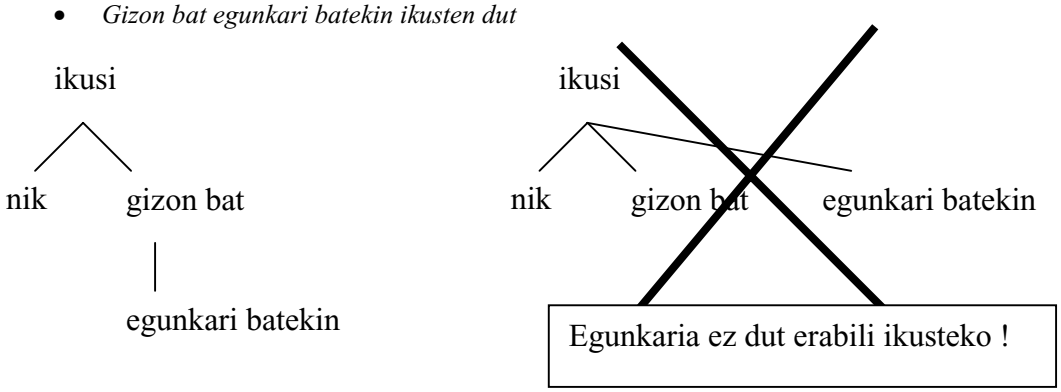
Informatikako programazio-lengoaia bateko *agindu* bat analizatu badugu, eta agindu horretan osagai bat bere kategoriako beste osagai batekin ordeztuz gero, bere analisiaren emaitza berdina izango da, nahikoa izango da osagaia ordeztea. Adibidez, ondoko aginduan aldagai bati espresio matematiko bat esleitzen

zaio eta dagokion analisia ezkerreko zuhaitza da; baina agindu horrexetan sinuaren osagaia (SIN) bere kategoriakoa den beste osagai batekin (kosinuaren funtzioa: COS) ordezten badugu, agindu berriaren analisian osagai berriaren ordezte hutsa eginez lortuko dugu (eskuineko zuhaitza):

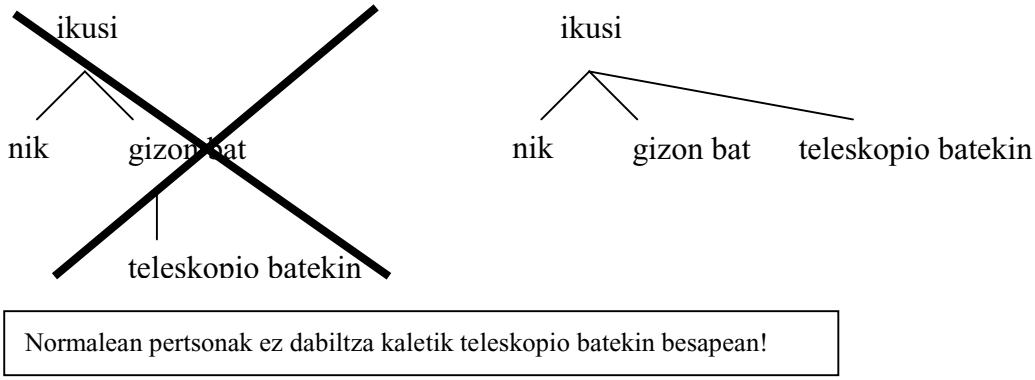
X := A * B + C * (D + (COS(Q) * SQRT (SIN(R)))

Hori informatikako programazio-lengoaietan gertatzen da, ez ordea hizkuntza tratatzerakoan. Esaldi bateko hitz bat aldatuz gero, esaldi berriaren analisia beti ez da izango esaldi zaharraren egitura berekoa. Hitz bat aldatuta esaldi osoaren egitura aldatzen da batzuetan, horrelakoetan ez da aldatzen hitzaren esanahia bakarrik, egitura ere aldatzen da. Adibidez, ondoko bi esaldiak oso antzekoak dira, izen bat bakarrik aldatzen da batetik bestera. Biek bina interpretazio sintaktiko eduki litzakete, baina, batzuetan, pertsona batek berehala baztertzen du bietako bat:

- *Gizon bat egunkari batekin ikusten dut*



- *Gizon bat teleskopio batekin ikusten dut*



Kasu bakoitzean egin den interpretazio sintaktikoa diferentea da, nahiz eta bi esaldiak oso antzekoak izan. *Teleskopio* hitzak ikusteko tresna adierazten du eta *egunkari* hitzak ez. Hori jakitea ezinbestekoa da interpretazio sintaktiko egokia zein den jakiteko. Are gehiago, pentsa genezake beste egoera bat non gizonak teleskopioa alboan duen, behatoki astronomiko bateko argazki batean, adibidez; horrela, bigarren esaldiko lehen interpretazio sintaktikoa ere zuzena litzateke. Hau da, oso zaila da bereiztea ezagutza lexikala (*teleskopio* eta *egunkari* hitzak), ezagutza sintaktikoa, semantikoa (*teleskopioa,* ikusteko tresna da, *egunkaria* ez) eta munduari buruzkoa (ea teleskopioa edo egunkaria besapean daraman edo alboan daukan). Lau ezagutza mota horiek batera erabili behar ditugu hizkuntzaren tratamendu automatikoan.

1.3 Bilakaera

Hastapenetan, lengoaia naturalaren prozesamenduaz (LNP) arduratzen zirenak (1950 eta 1960ko hamarkadetan), aplikazio zehatzetara mugatzen ziren batez ere, aplikaziotik aplikaziora helburuak aldatuz. Bi aplikazio multzo nagusi nabarmendu izan dira ordudanik:

1. Gizakien eta ordenagailuaren arteko komunikazioa errazten dutenak :
    - datu-baseen galdeketa-sistemak
    - elkarrizketarako interfazeak
2. Giza komunikaziorako aplikazioak :
    - testuen eduki-araketa
    - testu-edizioa
    - itzulpen automatikoa
    - hizketaren ezagutza eta sorkuntza

Garai haietatik hizkuntzaren teknologiak mota askotariko sistema informatikoak garatu nahi izan ditu, baina arlo horiek duela gutxi arte, batez ere, aplikazioetara lerratuta zeuden.

Hasieran sistema konputazional gehienek jostailuzko lexikoiak lantzen zituzten, oso aplikazio-domeinu konkretuei lotuak eta sarrera kopuru murritzekoak. Askotan zerrenda soilak baino ez ziren izaten. (Boguraev eta Briscoe, 1989:1)-n esaterako, hau diote:

> *"Knowledge of words underlies these tasks, yet until very recently dictionaries (or lexicons, as linguists usually call them) for natural language processing systems have by and large been the poor sisters of computational linguistic research".*

Bestalde, oro har, teoria linguistikoek sintaxi eta erregela gramatikaletan jartzen zituzten beren indarrak. 70 eta 80ko hamarkadetan LNPrekiko interesa areagotzeaz gain, azpimarratzekoa da epe horretan hurbilpen-aldaketa gertatu zela. Hau da, alderdi linguistikoan arreta handiagoa jarri zen. Hasieran, alderdi linguistikoak ez zuen garrantzi handirik, eta arestian aipatu dugun legez, aplikazioetara lerraturik zegoen hizkuntzalaritza konputazionala. Horrela, bada, garaturiko hainbat sistema aplikazio espezifikoetarako baino ez ziren baliagarri. Horren ondoren garaturiko beste bi joera nagusi ere aipatu beharrekoak dira. Batean, ordenagailuaz baliatuko dira modelo linguistiko teorikoak frogatzeko. Hots, teoriak sorrarazitako sistemak ditugu, eta garatu izan dira zenbait teoria frogatzeko; beste batzuen artean: gramatika transformazionalak (*Transformational Grammars*) (Friedman, 1969), Montague-ren gramatikak (Friedman, 1978), *Generalized Phrase Structure Grammars* (GPSG) (Evans, 1985; Phillips eta Thompson, 1985). Ikuspegi hau lantzen duten ikertzaileek tresna konputazionalak erabiltzen dituzte honakoa egiaztatu ahal izateko: ea proposaturiko eredu gramatikalak benetan sortzen dituen sortu beharko lituzkeen esaldiak . Ondorengoan, joera nagusia (egun ere dirauena) corpusetan oinarritzean datza. Horien artean ere ikuspegi ugari aurki ditzakegu, jakina, baina denak bat datoz hizkuntza aztertzeko corpusak ordenagailuez baliatuz ikertzerakoan.

Ondorengo urteetan, zenbait faktoreren eraginak ingeniaritza linguistikoko baliabideak eta aplikazioak ezinbesteko bilakatu ditu informazioaren gizartean. PCen agerpenak, haien kostua jaisteak eta prestazioak

hobetzeak (memoria eta prozesadorea) informatika eskuragarriago egin dute hainbat erabiltzailerentzat. Horrek ekarri du baliabide linguistikoen eskaeraren hazkundea, eta horrekin batera dokumentuen edizioan laguntzen duten testu-prozesadoreak. Nagusiki tresna hauek jokatuko dute paper nagusia: zuzentzaile ortografikoek, estilo-zuzentzaileek, sintaxi-zuzentzaileek eta sinonimoen hiztegiek. Baina, testu-edizioko tresna horiek baino askoz ere laguntza hobeak daude merkatuan eskuragarri, eta are laguntza bereziagoak bilatzen dira ikertokietan. Ordenagailuaren bitartez hizkuntzaren tratamendua egiten duten aplikazioak eta programak gero eta gehiago dira, ordenagailuarekiko komunikazioa egunero erabiltzen dugun hizkuntzaren bitartez egin ahal izatea gero eta normalagoa baita. Beste alde batetik, gizarte eleaniztunek hizkuntza diferenteen artean egin behar izaten dituzten joan-etorriak leuntzeko ere aparteko laguna izango dugu ordenagailua. Gainera, telekomunikazioetan gertatutako aurrerapen izugarriak eragin duen Internet fenomenoak izugarri areagotu egin du hizkuntzaren tratamendu automatikoaren beharra. Izan ere, nahiz eta informazio kopuru izugarria lortu ahal izan, ez da erraza bilatzen dugun informazioa aurkitzea, eta informazioa ondo selekzionatzeko tratamendu linguistikoa lagungarria baino areago ezinbestekoa da.

Gaur egungo joera, ordea, hastapenetakoarekin alderatuz gero, erabat aldatu dela esan dezakegu. Hizkuntzalaritza teorikoaren zein HKren egungo joeraren arabera, hizkuntza-ezagutza gramatikaren arlotik lexikoarenera lerratzen baita. Teoria linguistikoan eragin handiena izan duten formalismoek (UG, LFG, HPSG... aurrerago azalduko ditugunak dagokien atalean) erregela gramatikalak erraztera jotzen dute, eta lexikoa muina izango dute. Alderdi teorikoari dagokionez, segur aski, Chomsky-k eman zion abiada joera horri (Chomsky, 1970). Ildo beretik jardungo dute aplikazioei loturiko LNPko arlokoek ere. Hau da, sistema errealetarako ezagutza lexikala eskuratzea ezinbestekotzat jotzen da laborategiko saioak gaindituz arlo honetan aurrera egin nahi bada. LNPko sistemek neurri errealeko osagai lexikalak behar dituzte, aplikazio-eremua hedatu eta sendotzeko. Baina osagai lexikal horiek eskuz egitea hain da lan handia, ezen ezinezkoa baita ia. Horrela, bada, LNPko aplikazioen problemarik larriena lexikoi konputazionalak hornitzeko ezagutza lexikalaren eskuratze-prozesuak garatzean datza. Gauzak horrela, LNPrako lexikoien eraikuntzarako laguntza automatikoak garatzea eta dauden baliabide lexikalez baliatzea dira harturiko irtenbide nagusiak.

Bestalde, hizkuntzen teknologiako aplikazioak diseinatzerakoan, ikertzaile asko datu estatistikoek gidaturiko metodoetara lerratu da azken hamarkadan. Zenbait hamarkadatan, kognizio-egiturak eta giza hizkuntzaren erabiltzailearen prozesuen azterketatik teknologiak aurrera egin zezakeelako itxaropena izan ondoren, gizakiek sorturiko datu linguistikoetan eta hizkuntzaren teknologiak prozesatu beharrekoetan jarri dute ikusmira.

Dena dela, hizkuntzaren tratamendu automatikoaren arloan izandako lorpenek muga handiak dituzte oraindik, hizkuntza ulertzea eta sortzea oso zaila baita. Baina horrek ez du esan nahi hizkuntza lantzeko tresna automatikoak utopia direnik, hizkuntzaren oinarrizko ezagutza minimo batekin laguntza interesgarriak eskain daitezke eta. Emaitza probetxugarriak lortzeko, ordenagailuaren lana aztergai espezifiko eta mugatu batean kokatu behar da. Egun aurretiko hitzordua ematen duten sistema gehienek zenbakiak eta astegunen izenak besterik ez dute ulertzen, baina hala ere ekonomikoki oso interesgarriak diren aplikazioak antolatu dira horrekin. Etorkizunean, aplikazio mugatuko sistemak bilduz, ahalmen

handiagoko sistema berriak lor litezke, baina egun ibili dabiltzan aplikazioek helburu espezifikoak dituzte.

1.4 LNPko sistema batean landu behar diren ezagutza motak[2]

LNPko sistema batek hizkuntzaren tratamendu osoa egin behar badu, honako ezagutza mota hauek erabili beharko ditu:

- **Fonetikoa eta fonologikoa.** Zehazten dute nola ahoskatu behar diren hitzak eta letra bakoitzari zein fonema dagokion. Hizketaren tratamenduaz aritzerakoan, bi sistema nagusi garatzen dira: hizketaren ezagumendua edo analisia (*Speech Recognition,* SR), eta sintesia edo sorkuntza.
- **Lexikala.** Hizkuntzan erabil daitezkeen morfemak zehazten dira hemen (lemak, aurrizkiak, artizkiak eta atzizkiak), eta bakoitzarentzat bere hizkuntza-ezaugarriak zehazten dira.
- **Morfologikoa.** Hitz posibleen osaketa definitzen da morfemak erabiliz. Zein morfema-kate dira posible eta zeintzuk ez? Morfema pare bat biltzen direnean letrarik galtzen da? edo gehitu behar da? edo aldatu? Aplikazio batzuetan ez da beharrezkoa, adibidez, ingeleserako askotan ez da kontuan hartua izaten (hala ere, bada bestela pentsatzen duenik ere). Baina, morfologia aberatsa duten hizkuntzen prozesamenduan oso garrantzitsua da, esate baterako, euskara, suomiera, etab.
- **Sintaktikoa.** Esaldien egitura ezagutzeaz arduratzen da, hau da, hitz bakoitza zeinekin datorren. Hitzen arteko harremanak definitzen dira hemen, haien kategoria sintaktikoen arabera.
- **Semantikoa.** Hitzen esanahia lortu eta hitzen esanahietatik abiatuz, beraiek osatzen duten esaldiaren esanahia lortu.
- **Testuinguruari dagokiona.** Pragmatika gisa ezagutzen dena. Berez linguistikoa ez den, eta igorpen linguistikoen prozesamenduan eta interpretazioan eragina duten informazioez arduratzen da. Bi atal bereiz daitezke:
    - **Diskurtsoaren ezagutza.** Hizkuntza erabiliz komunikatzeko gizakioi suposatzen zaigun ezagutza. Lehenago igorri diren esaldien interpretazioak kontuan hartzen dira izenorde, elipsi eta denbora-aspektuak egoki ulertu ahal izateko. Hizketako parte-hartzaile bakoitzak besteek dakitenari buruz edo nahi dutenari buruz suposatzen duena ere jakin behar da, elkarrizketa bat konputagailu bidez eraman ahal izateko.
    - **Munduaren gaineko ezagutza.** Hizkuntza bateko hiztunek elkarren artean komunikatzerakoan, munduari buruz duten ezagutza kontzeptual guztia hartu behar da kontuan, alegia, mintzagaiari berari buruz jakin behar dena. Horrelako ezagutzak esaldietan esplizituki adierazten ez den eta bistan den informazioa ulertzeko balio du.

Jakina, goian aipatu ditugun modulu horietako informazio guztia ez da beharrezkoa lexikoi espezifiko

---

2 Joerak azkar ari dira aldatzen, eta jasotzen diren atalak tradizionalki jasotzen direnak dira. Bereziki, atal hau lantzeko ondoko liburu hau baliatu dugu: *Survey of the State of the Art in Human Language Technology*, Edited by Ronald Cole, Joseph Mariani, Hans Uszkoreit, Giovanni Batista Varile, Annie Zaenen eta Antonio Zampolli.

baterako. Lexikoi bakoitzak LNPko sistema espezifiko baterako beharren araberako informazioa izango du eta. Hain zuzen, liburuaren atal nagusi bat osatuz, 2. kapitulutik 5.era ezagutza mota hauetan landu beharreko informazioan sakonduko dugu. Esan behar da, hala ere, fonetika/fonologiaren alorrak ez ditugula garatuko, horrekin loturiko oinarri, tresna eta aplikazioak gutxi direlako. Horretaz gain, ahozkoarekin lotuago dago, eta idatzizkoaren gainean lan egiten dutenek ez dute horretan lan egiten, eta orobat alderantziz. Alor hauek besteetatik nahiko bereizita ikertzen direla esan daiteke.

Joerak azkar ari dira aldatzen, eta aipatzen diren atalak tradizionalki landu direnak dira. Guk, atal hau lantzeko, bereziki Cole *et al.* (1998) liburua erabili dugu. Autore askoren artean idatzitako liburu honetan LNPri buruzko ikuspegi zabala jasotzen da.

1.5 Analisiaren faseak eta beraien arteko koordinazioa

Beti horrela egiten ez bada ere, argigarria izaten da hiru fase nagusi identifikatzea esaldien tratamendu automatikoan: 1) analisi sintaktikoa, 2) interpretazio semantikoa, eta 3) testuinguruko interpretazioa.

Fase bakoitzean zenbait ezagutza mota erabiltzen dira. Atal honen bukaeran ikusiko dugunez, sistema guztietan ez dira errespetatzen fase hauek, eta gainera, beti ez dira egiten linealki bata bestearen atzean. Esaldien ulerkuntza egin behar denean, ordena horretan burutu ohi dira, eta, esaldien sorkuntza egin behar denean, alderantzizkoan.

Fase bakoitzaren zeregina deskribatzeko sarrera (zer datu hartzen du?) eta irteera (zer lortzen du? zein da emaitza?) deskribatuko ditugu hasieran; gero, adibide batzuk eman, eta bukaeran, saiatuko gara definitzen fase horretan nola mozten den anbiguotasuna.

1.5.1 Analisi sintaktikoa (lexiko-morfosintaktikoa)

Analisi sintaktikoa esaten dugu hemen, baina zuzenagoa litzateke analisi lexiko-morfosintaktikoa esatea.

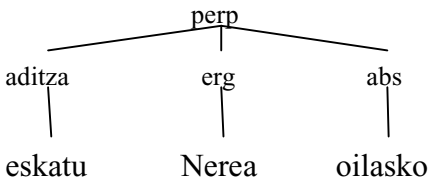
Sarrera: Esaldi bat.
Irteera: Hitzen arteko erlazio sintaktikoak adierazten dituen egitura. Emaitza esaldi osorako egitura bakarra izan daiteke (gehienetan zuhaitza), edo esaldi osorako zenbait egitura baldin zenbait interpretazio sintaktiko posible onartzen baditu. Ez bada lortzen esaldi osorako analisirik; hau da, esaldia gramatikala ez bada (gure gramatikaren arabera, noski), esaldi barruko zati zuzen handienen edo osagai sintaktikoen lista lortzen da.

1. adibidea: Esaldia hau bada: *Nereak oilaskoa eskatu du,*

emaitza hauetako bat izan daiteke.

```
(perp (erg (is (izb nerea (num s) (mug m)))
       (abs (is (ize oilasko (num s) (mug m)))
       (aditza (ad eskatu) (aldi lehen) ...
   ...)
```

Cambridge University Press. 1998. Autore askoren artean idatzitako liburu honetan LNPri buruzko ikuspegi zabala jasotzen da.



2. adibidea: Esaldia hau bada: **Beltz etorri zenbait gizon dira*,

esaldi ezegoki hori analizatuz gero, ez da analisi osorik lortuko eta zati hauek eskainiko dira emaitza gisa:

(Beltz) (etorri) (zenbait gizon) (dira)

Anbiguotasunaren murrizketa: Anbiguotasun lexikala eta morfologikoa mozten saiatzen da.

- Hitzen analisi morfologikoa (morfosintaktikoa) egiten da fase honen barruan.
- Hitzaren analisi bat baztertu egingo da lexikoian lema ez badago (*kerekere), bateraezinak diren morfemak lotuz osatu bada (**zenbaitgo*= zenbait+ago), edo bateratzean egin beharreko letra-aldaketak ondo egin ez badira (**amaek*=ama+ek).
- Inguruko hitzen artean egoki ez diren interpretazio morfosintaktikoak baztertu egiten dira. Adibidez "*ama gorri hori duen katuak*" perpauseko hitzek bakoitzak bere aldetik, modu isolatuan, izan ditzakeen interpretazioen arteko zenbait ez dira posible testuinguru horretan:

ama	**gorri**	**hori**	**zuen**	**katuak**
~~ADI~~	~~ADI~~	~~ADI~~	~~ADL(3)~~	IZE(2)
IZE(3)	ADJ(2)	~~ADJ(2)~~	ADT	
	~~IZE~~	DET	~~IOR(4)~~	

1.5.2 Interpretazio semantikoa

Sarrera: Esaldi baten egitura sintaktikoa.

Irteera: Esaldiaren esanahiaren errepresentazioa, baina esanahi abstraktua, testuingurua oraindik kontuan hartu gabe. Kontuan hartzeko elementuak hauexek lirateke:

- Lexikoian jar daitekeen informazio semantikoa (hitzen esanahia).
- Hitz batek hainbat esanahi eduki ditzakeela.
- Esaldiaren esanahiaren errepresentazioa eraikitzen da bere osagaien interpretazioekin.

Adibidea: *Nereak oilaskoa eskatu du* esaldiaren analisi sintaktikoa honako hau bazen:

```
(perp (erg (is (izb nerea (num s) (mug m)))
       (abs (is (ize oilasko (num s) (mug m)))
       (ad (ad eskatu) (aldi lehen) ...
      ...)
```

Analisi hori datu gisa interpretazio semantikora eramanda emaitza hau litzateke:

```
(ESKATU4 ?V   (AGENTE (IZB ?Y (IZB NEREA1))
              (TEMA    (MS ?Z OILASKO2)))
```

non ezagutu diren *eskatu, oilasko* eta *Nerea* hitzen adiera egokiak (eskatu4, oilasko2 eta Nerea1),[3] non *eskatu4* aditzaren kasu semantikoak identifikatu diren eta semantikoki posibleak

[3] Eskatu4: eskatu aditzaren 4. adiera da, oilasko2: oilasko hitzaren 2. adiera, ...

direla egiaztatu den (*eskatu4* ekintzetan agentea pertsona bat izan behar da; gure esaldian *Nerea1* pertsona da. Bestalde, *eskatu4* ekintzetan tema janari mota bat izan behar da eta *Oilasko2* semantikoki janaria da).

Anbiguotasuna murrizten:

- Baztertzen dira esanahi ulergarria ez duten egitura sintaktikoak. Adibidez:

  **Asmo berde kolorebakoak bortizki lo egiten zuten,*

  baztertuko genuke ondoko arrazoi semantiko hauengatik:

  - asmoek kolorerik ez dutelako, ez baitira objektu fisikoak
  - berdea eta kolorebakoa ez dira bateragarriak
  - ezin da bortizki lo egin
  - asmoek ezin dute lo egin, ez baitira izaki bizidunak

- Baztertzen dira hitzen esanahi ezegokiak. Adibidez perpaus honetan ba al dago *zubia* hitzaren esanahiren bat baztertzea? Posibleak al dira *Zubia-eraikuntza* eta *zubia-asteburu luzea* adiera biak testuinguru honetan?

  Londresera joan naiz abenduko zubian

  Kasu honetan, zubia asteburu luze bat bezala ulertu behar da.

1.5.3 Testuinguruko interpretazioa

Sarrera: Esaldiaren esanahiaren errepresentazioa.

Aurreko esaldien interpretazioak.

Eta ezagutza hizketa-gaiari buruzkoa (munduari buruzkoa) eta hizketari buruzkoa.

Irteera:

- Esaldiaren esanahi konkretuaren errepresentazioa (testuingurua kontuan hartuta).
- Aurreko esaldien interpretazioak kontuan hartuta, izenorde, elipsi eta denbora-aspektuak era egokian osatu dira.
- Esaldian esplizituki adierazten ez dena, baina munduari buruzko ezagutza edo diskurtsoari buruzkoa erabiliz zentzuzkoa dena, esplizitu egin da azken interpretazio honetan

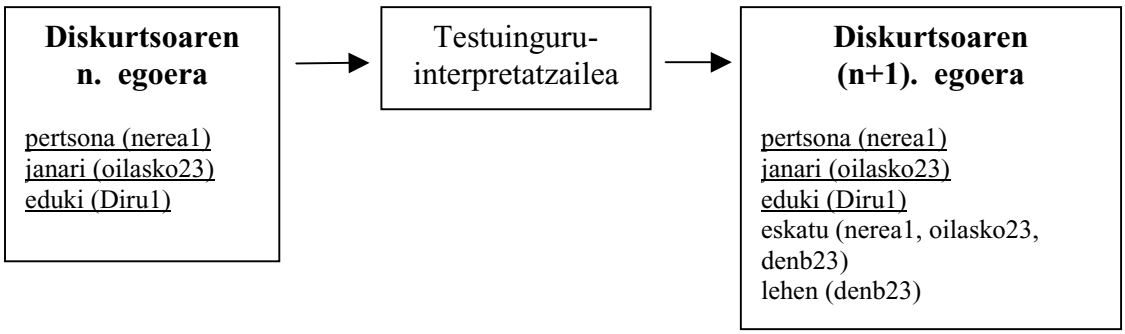
Adibidea: *Nereak oilaskoa eskatu du* esaldiaren interpretazio sintaktikoa honako hau bazen:

```
(perp (erg (is (izb nerea (num s) (mug m)))
       (abs (is (ize oilasko (num s) (mug m)))
       (ad (ad eskatu) (aldi lehen) ...
   ...)
```

eta bere interpretazio semantikoa hau:

```
(ESKATU1 ?V      (AGENTE (IZB ?Y (IZB NEREA1))
                 (TEMA    (MS ?Z OILASKO2)))
```

testuinguruko interpretazioaren emaitza berbaldia edo diskurtsoaren egoera berri bat sortzea litzateke, non aurreko egoerari informazio berria gehitu zaion. Gure adibidean azkeneko bi klausula gehitu dira: *eskatu (nerea1, oilasko23, denb23)* eta *lehen (denb23).*



2. irudia. Diskurtsoan dauden egoeren errepresentazioa

Anbiguotasuna murrizten:

- Baztertzen dira egoera konkretu horretan zentzurik ez duten interpretazio semantikoak. Batzuetan esaldiaren esanahia berrinterpretatu behar da testuinguruko informazioarekin osatuz. Adibidez, *Badakizu zer ordu den?* galderari erantzun desberdinak emango dizkiogu testuinguru desberdinetan, egiten dugun interpretazioa oso bestelakoa baita:

  –*Zazpi t'erdiak*
  Kalean noala erlojurik ez daukan neskato batek galdetzen badit.
  –*Bai*
  Kalean noala erlojurik ez daukan lagun adarjotzaileak galdetzen badit.
  –*Autobusa galdu dut*
  Zinemako atean sarrerak eskuan zituela ordubetez nire zain egon den lagunari.

  Adibidez, lehen esan dugu *bosgarren udako unibertsitatea* anbiguoa dela, baina testuinguru konkretu batean baldin badakigu udako unibertsitate hori behin bakarrik antolatu dela eta gehiagotan ez dela antolatuko (munduko ezagutza), argi dago testuinguru horretan bigarren interpretazioak ez duela zentzurik eta ken dezakegula:

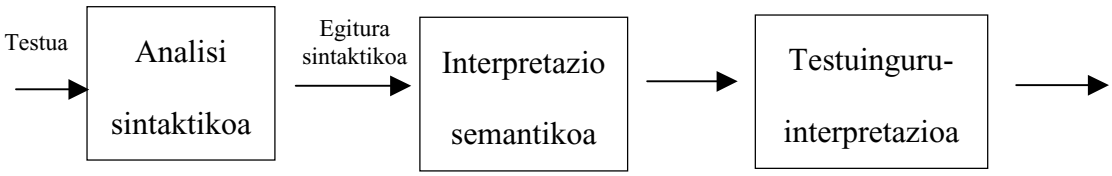
  - *(Bosgarren udako) unibertsitatea*. Bosgarren udan antolatu zen unibertsitatea
  - ~~*Bostgarren (udako unibertsitatea)*~~. Bosgarren aldian antolatu zen udako unibertsitatea

1.5.4 Faseen arteko koordinazioa

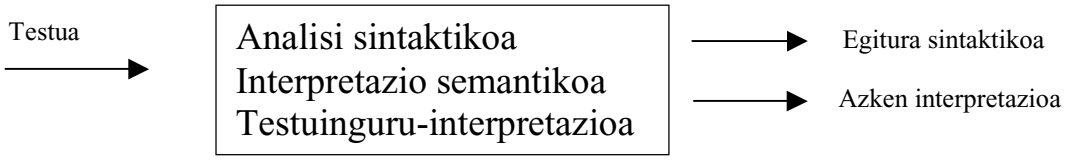
Definitu ditugun hiru fase horiek beti ez dira egikaritzen edo egiten bata bestearen atzetik sekuentzialki. Klasikoki bai, baina badira alternatibak, eta gainera egungo LNPko sistema batzuetan, paradigma estokastikoa erabiltzen duten sistemetan fase horiek ez dira inondik ere agertzen kasu honetan, ezagutza linguistikoa zeharo baztertzen baitute.

Hori aurretik argituta, hemen hiru aukera nagusi aurkeztuko ditugu fase horiek koordinatzeko:

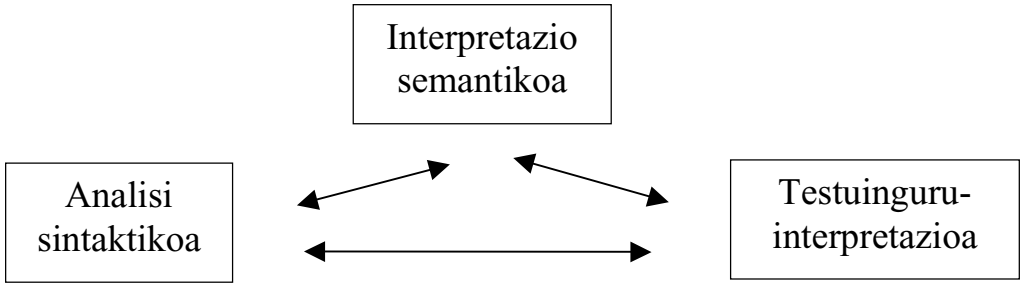
A) Faseen egikaritze sekuentziala. Faseen adibideekin suposatzen genuena, fase berri bat ez da bukatzen aurrekoa guztiz bukatu arte. Fase bakoitzean erabiltzen diren ezagutzak eta tresnak zeharo independenteak dira.



B) Faseen egikaritze paraleloa. Fase guztiak paraleloan egikaritzen dira. Ezagutza linguistikoak (sintaktiko, semantiko eta pragmatikoa) eta tresnak ez daude bereizita. Modulu horiek elkarren artean lan egiten dute.



C) Faseen egikaritze modularra. Fase guztiak paraleloan egikaritzen dira. Ezagutza linguistikoak (sintaktiko, semantiko eta pragmatikoa) eta moduluak bereizita daude. Modulu horiek elkarren artean lan egiten dute.



1.6 Hizkuntza-teknologiako produktuak

Hizkuntza-teknologiako produktuetan aplikagarritasun maila hauek bereizten ditugu:

- Aplikazioak: erabiltzaile arruntarentzat salgai diren programak.
- Tresnak: hizkuntza-teknologiako ekoizleentzat bakarrik interesgarriak direnak, produktu berriak garatzeko baliagarriak.
- Oinarriak: ikerketarako edota edozein aplikazio edo tresna garatzeko behar-beharrezkoak diren oinarri linguistikoak.

Produktuei buruzko berri zehatzagoa liburu honen hirugarren atal nagusian (produktuak) emango dugu. Hemen, gainbegiratu baino ez dugu egingo.

1.6.1 Erabiltzaile arruntentzako aplikazioak

Modu eskematiko batez, hauek dira erabiltzaile arruntentzako zenbait produktu (Alegria *et al.,* 1997):

- Testuen edizioa eta kudeaketa. Egun, badira testu-egileari eskaintzen zaizkion laguntza bereziak. Ikus ditzagun orain zein diren garrantzitsuenak.
    - Ortografia-zuzentzaileek bete dituzte urte batzuk merkatuan, eta gaur egun hizkuntza askotarako aurki daitezke. Euskarari dagokionez, 1994tik dago dendetan XUXEN euskararako egiaztatzaile/zuzentzaile ortografikoa.
    - Idazkera- eta sintaxi-zuzentzaileak ere merkaturatu dira zenbait hizkuntzarentzat. Testuingurua kontuan hartzen dute eta, adibidez, "nik joan naiz" esaldia prozesatuz gero, ortografia-zuzentzaileak ez luke errorerik salatuko, hiru hitzok isolatuta posible baitira, baina sintaxi-zuzentzaileak testuinguru horretan "nik" hitza gaizki dagoela salatuko luke eta "ni" izan beharko lukeela proposatu.
    - Laguntza lexikaletan edozein hitzen sinonimo edo antonimoak lor daitezke testu-prozesaketako programatik atera gabe, baita taxonomikoki konkretuagoak edo orokorragoak diren antzeko hitzak ere (adibidez: intsektu hitzetik orokorragoa den animalia edo konkretuagoak diren iñurri, euli...), thesaurusa kontsultatuz.
    - Testu eleaniztunak lantzeko, prozesadore zabalduenetan zenbait programa integratzen dira. Programa horietan, glosategi, hiztegi eta itzulpenen berrerabilerarako laguntzak eskaintzen dira. Adibide gisa Siemens-en Eurolang Optimizer, IBMren TranslationManager/2 eta Trados-en Translation Workbench programak ditugu.
- Testu-masa handiak tratatzeko edo kudeatzeko aplikazio nagusiak lau dira:
    - Kontzeptu-bilatzaileak. On-line moduko kontzeptu-bilatzaileen inguruan mila milioi dolarreko industria antolatuta zegoen 1994an. Euskararako ere bada berriki Ametzagaina taldeak kaleratutako Kapsula softwarea, euskarazko dokumentu-baseen kudeaketara zuzendua.
    - Kategorizazio-sistemak oso baliagarriak dira makina bat dokumentu (adibidez: telefonoetako matxura-parteak, albisteak, hildako militarren parteak, marketineko datuak...) kategoria multzo txiki baten arabera sailkatu behar izanez gero.
    - Informazio-erauzketarako sistemek lengoaia naturalez idatziriko testuetatik datu-base egituratu bat osatzen dute. Azken helburua albiste multzo handi batetik abiatuz fitxa konkretuak betetzea litzateke, nork-nori-zer egin dion jakiteko.
    - Testu-sorkuntza automatikoa informazio-erauzketaren kontrakoa da. Kasu honetan ordenagailu barruan dauden datu konplexuetatik abiatuta (inprimakiak, datu kodetuak edo zenbakizko formatuan dauden informazioak...), datu horien edukia azalduko zaio erabiltzaileari bere hizkuntzan.
- Itzulpen automatikoa. Produktu ugari dago merkatuan salgai testu-itzulpenean laguntza emateko, baina euskara tratatzen duen sistemarik ez dago. Itzulpenaren automatizazioa ez da inoiz erabatekoa, eta automatizazio mailaren arabera ondoko sailkapena egiten da: 1) erabateko

itzulpen automatikoa; 2) giza laguntzaz buruturiko ordenagailu bidezko itzulpena; 3) ordenagailuz lagunduriko giza itzulpena; 4) datu-banku terminologikoak.

- Ordenagailuen erabilera lengoaia naturalaren bidez. Aplikazio mota honetako sistemek ordenagailuaren eta gizakiaren arteko komunikazioa errazten dute, erabiltzaileek bere hizkuntzaz lan egiteko aukera du eta. Helburu orokorrekorik ez da luzaroan salgai egongo, baina badira dagoeneko aplikazio konkretuei lotuta dauden batzuk. Datu-baseetarako galdeketa-sistema ugari dago, batez ere ingelesez.
- Ahozko hizkuntzaren tratamendua. Sistema gehienek oso hitz gutxi ezagutzen dituzte, eta horien artean beti daude zenbakiak. Beste alde batetik, gero eta arruntago bihurtzen ari zaigu makinen ahots sintetizatuak entzutea gasolindegietan edo tabako-edariak saltzen dituzten makinetan. Ahozko hizkuntzaren tratamenduko teknikak antzeko beste aplikazioetan ere erabiltzen dira: eskuz idatzitako testuak ezagutzeko edota testu elektronikoen bertsio elektronikoa lortzen duten OCR (*Optical Character Recognizer,* karaktere-ezagutzaile optikoak) izenekoetan.

1.6.2 Tresnak

Atal honetan hizkuntzaren tratamendurako aplikazio-ekoizleentzat edo arloko ikertzaileentzat interesgarriak diren tresnak aipatuko ditugu. Tresna horiek ez daude diseinaturik, oro har, erabiltzaile arruntarentzat.

- Analizatzaile morfologikoak. Hizkuntza flexionatu eta eranskarien kasuan –hala nola euskara– ezinbestekoak dira ondorengo aplikazioetarako:
    - Zuzentzaile ortografikoa.
    - Tutore-sistema automatikoak hizkuntza ikasten ari den jendearentzat.
    - OCR dokumentuen irakurketan (eskanerrak erabiltzean) sor daitezkeen erroreak detektatzeko.
    - Hizketaren sintesia edo testu-sorkuntza lortzeko sorkuntza morfologikoa funtsezko osagarria da.
    - Hizkuntza-aplikazio sofistikatuagoetarako –sintaxian oinarritutakoak, itzulpen automatikoa, etab.— lehen urrats gisa.
- Lematizatzaile/etiketatzaileak. Etiketatzaileek testuko hitz bakoitzak dituen analisi guztien artean zuzena dena aukeratu behar dute; lematizatzaileek, aldiz, lema posibleen artean dagokiona. Tresna hauek izan duten arrakasta beren aplikazioetan datza, oso aplikazio interesgarri eta aktualak baitituzte:
    - Indexazioa: testuak indexatu nahi direnean ez zaigu forma interesatzen, lema eta kategoria baizik. Indexazioa da oinarria gaur egun hain modan dauden datu-base dokumentaletan eta Interneteko bilatzaileetan. Adibidez, testu batean *kalekoak, kalera* eta *kalejiratik* agertzen badira, lehen biek azaldu behar dute *kaleaz* galdetzen dugunean, baina hirugarrenak *kalejiraz* egiten dugunean.

- Terminologia/lexikografia: automatikoki lemak ondo identifikatzen badira eta dagozkien etiketak egokitzen bazaizkie lan lexikografikoa erruz errazten da, eta testu batetik terminologia automatikoki erauzteak ez dirudi oso lan zaila.

- Analizatzaile sintaktikoak. Analizatzaile sintaktikoen zeregina testuetako osagai sintaktikoak ezagutzea da: hitz isolatuez osatu sekuentzietan elkarri lotuta dauden egitura sintaktikoak (perpausak, izen-sintagmak, aditz-sintagmak, izenlagunak, eta abar) ezagutuko dira.

1.6.3 Baliabide linguistikoak

Atal hau euskarari lotutako baliabide linguistikoen bidez azalduko dugu, eta gehienbat IXA taldean garatu diren baliabide linguistikoen bidez. Izatez, IXA taldeak euskararako garatu dituenak dira hizkuntzaren tratamendu automatikoan, oro har, aurkitzen ditugunak:

- Datu-base lexikala eta morfologiaren deskribapena. Datu-base lexikala da hizkuntzaren lexikoaren biltegi erraldoia. Hiztegi elektronikoen moduko bat da, hizkuntzaren tratamendu automatikoari begira eraikia, eta, beraz, hizkuntzaren tratamendua automatizatu nahi horrek dituen eskakizunak kontuan harturik antolatua. EDBL, Euskararen Datu-Base Lexikala dugu IXA taldeak garatutako oinarri lexikala, etengabe eguneratuz doana, eta gaur edo bihar komunitate zabalago bati bere ateak irekiko dizkiona, oinarriak prestatze-bide honetaz beste batzuk ere balia daitezen.
- Hiztegi elektronikoak. Hizkuntzaren datu-base lexikal orokorra oinarri dela, horren inguruan biltzen ahal dira beste zenbait tresna lexikal ere: definizio-hiztegiak, hiztegi terminologiko berezituak, hiztegi elebidunak, eta beste. Hor ditugu UZEIren Euskalterm datu-banku terminologikoa, Sinonimoen hiztegia eta Atzekoz aurrera (hitz-bukaeren hiztegia); I. Sarasolaren *Euskal Hiztegia;* eta Elhuyarrek, Harluxet Fundazioak eta Adorez taldeak, besteak beste, euskarri elektronikoan kaleratutako hiztegi-lanak .
- Gramatika konputazionalak. Bi hurbilpen desberdinetatik, euskararako bi gramatika konputazional landu dira IXA taldean:
  - PATR-II izeneko baterakuntza-formalismoaz (Shieber, 1987) baliatuz. Izen-sintagma eta perpaus bakunen egitura deskribatzen duen gramatika.
  - Murriztapen Gramatika formalismoa (Karlsson *et al.,* 1995) baliatuz. Batez ere, desanbiguazio morfosintaktikorako erabili da.
- Taxonomia semantikoak. Kontzeptuen artean hainbat motatako harremanak ezarriz egiten diren sare semantikoak dira. Ingelesezko, Wordnet izenekoa (Miller, 1990) da sare semantiko ezagunena, eta hori abiapuntutzat hartuz, euskararako halako sarea (EuskalWordNet) eratzen dihardu IXA taldeak.
- Hizkuntza-corpusak. Hizkuntza-corpusak testu-masa handiak dira, informazio linguistikoaren iturri nagusietariko bat eta arestian aipatutako aplikazio, tresna eta oinarrietarako probaleku ezinbestekoak. Lexikografiarako bezalaxe, LNPrako ere ezinbestekoak ditugu hizkuntza-corpusak. Aipagarrienen artean dugu *XX. mendeko euskararen corpus estatistikoa*

(www.euskaracorpusa.net). Baita berezituagoa den *Ereduzko prosa gaur* ere (http://www.ehu.es/euskara-orria/euskara/ereduzkoa). Corpus hauek, ordea, ez dira nahikoak.

Beste hizkuntzetarako daudenekin konparatuta, eskas samar geratzen dira gure corpusak, bai tamaina aldetik bai testu barruan etiketatzen denaren aldetik (lemak, hitzen kategoria desanbiguatuak, osagai sintaktikoak, hitzen adiera desanbiguatuak...):

Corpusa	Hitz kopurua	Hizkuntza
British National Corpus	100 milioi hitz	Ingelesa
Bank of English (COBUILD)	300 milioi hitz	Ingelesa
FRANTEXT	150 milioi hitz	Frantsesa
CRAE	130 milioi hitz	Gaztelania
CORDE	136 milioi hitz	Gaztelania
XX. mendeko corpus estatistikoa	5 milioi hitz	Euskara

Testuak ondo aukeratuz gero, azterketaren emaitzak hizkuntzaren egoeraren adierazgarriak eta eredugarriak izan daitezke, erreferentzia estandarra hizkuntza lantzeko. Informazioaren gizartean, hizkuntza batek duen garrantzia neurtzeko garaian, aplikazioak garatzeko dituen baliabide linguistikoak aztertzen dira gaur egun. Baliabide horien artean, corpus handien garapena lehenetariko helburua izan ohi da.

Horregatik, euskarazko corpusen biltze-lan eta antolaketa sistematikoari ekin behar zaio lehenbailehen, modu planifikatu batean. Lan horretan toki askotako jendeak hartu behar luke parte –Euskaltzaindia, UZEI, komunikabideak, argitaletxeak, eta abar— uste baitugu halako lan bat behar-beharrezkoa dela, honetan ari garenontzat ez ezik, baita beste ikertzaile askorentzat ere.

Bigarren atala

LNP-KO SISTEMA BATEAN LANDU BEHAR DIREN EZAGUTZA MOTAK

1.4 puntuan LNPko sistema batean landu beharreko ezagutza motak aipatu ditugu. Hemen berorietan sakonduko dugu, hizkuntzalaritza konputazionalean dauden lan eta ikergaien ikuspegi orokorra emateko. Xehetasunak, etorkizunean alor bakoitzerako egitekoak diren liburukietan emango dira. Morfologiari dagokiona jada argitaratu da (Alegria & Urkiak, 2002).

Ezagutza mota guztiak aztertzerakoan, oro har, ondoko eskemari jarraituko diogu:

- Sarrera: definizioa eta helburuak.
- Oinarrizko kontzeptuak.
- Hurbilpenak. Ezagutza mota hori lantzeko sortu diren metodologiak, teknikak edo estrategiak.
- Egungo egoera.

Zenbait ezagutza mota konputazionalki sakonkiago aztertu dira, zailtasunak zailtasun, horietan oinarri eta tresna sendoak eraiki ahal izan baitira. Ondorioz, ezagutza mota bakoitzeko edukia ez da proportzionala izango. Zehazki, morfologia, sintaxia eta semantika sakonkiago aztertuko ditugu; pragmatikaz eta hizketaren tratamenduaz, berriz, zertzelada batzuk baino ez ditugu emango. Bestetik, esana dugun bezala, fonetika eta fonologia alorrak ez ditugu garatuko; izan ere, horiei loturiko oinarri, tresna eta aplikazioak gutxi dira oraindaino. Horretaz gain, ahozkoari lotuago daude, eta idatzizkoaren gainean lan egiten dutenek ez dute berorietan lan egiten, eta orobat alderantziz. Fonologia eta fonetika alorrak besteetatik nahiko bereizita ikertzen direla esan daiteke.

Honenbestez, lau dira aztertuko ditugun ezagutza motak:

- Morfologia
- Sintaxia
- Semantika
- Pragmatika

2 Morfologia

2.1 Sarrera

Hizkuntza baten tratamendu automatikoan, lehenengo urratsa hitzen osaera morfologikoa aztertzea izan ohi da, hau da, hitz bat osatzen duten zati morfologikoak bereiztea eta, era berean, zati horiek bata bestearekin lotzeko arauak definitzea[4].

Horretarako, lehenik eta behin *hitz* kontzeptuak ikuspegi konputazionalean hartzen duen zentzua argitu behar dugu. Hitza, zuriunetik zuriunerako karaktere-segida bat da, eta esaldia zuriunez bereizitako hitzez osatuta dago (Fontenelle *et al.,* 1994).

Hizkuntza guztiak ez dira morfologia-eredu berekoak: zenbaitzuk morfologia sinpleagokoak dira, eta beste batzuk konplexuagokoak (*the one in the mountains* vs *mendietakoa*). Ingelesez, esaterako, *mountain* lema *mountain* eta *mountains* hitz bezala (singularra eta plurala adieraziz, alegia) ager daiteke bakarrik; euskaraz, berriz, *mendi* lemak *mendia*, *mendiak*, *mendietan*, *mendietakoa* eta beste milaka hitz-forma osa ditzake. Horrela izanik, ingelesezko lexikoi konputazional bat osatzeko, ez da hainbesteko lana *mountain* eta *mountains* hitzei sarrera ematea; hots, aparteko bi hitz lantzea. Aldiz, hori pentsaezina da euskara gisako hizkuntza batean. Lemaren eta atzizkien arteko konbinazio guztiak kontuan izanez gero, lema bakoitzeko milaka eta milaka sarrera behar genituzke.

Bestalde, hitza osatzen duten morfema guztiak ez dira izaera berekoak: morfema batzuek hitzaren lema osatzen dute, alegia, hitzari esanahia ematen diote (pilota+*gile;* pilota-*partida*), eta beste batzuek flexioa osatzen dute (pilota +*a*+ *k*), hots, hitzari informazio morfosintaktikoa ematen diote.

Zenbait hizkuntzak informazio mota hori garbiki eta independenteki bereizten dute morfemaka (mendi + *ko* + *e* + *n*) , eta beste batzuek, ez (rosa +*ae*). Lehenengoei *hizkuntza eranskari* izena ematen zaie, eta bigarrenei *hizkuntza flexibo.*

Ikusi ditugun morfema zatiak elkarren artean ez dira edonola lotzen. Batzuetan aldaketa morfofonologikoak gertatzen dira (zakur + a –> zakurRa; aberats + tu –> abera(T)stu), eta horiek ere zehaztu egin behar dira.

Hitzen morfologia aztertzea, honenbestez, ez da gauza sinplea.

Ingelesa aztertzearen ondorioz, eta hizkuntza honek ez duenez morfologia konplexua, morfologiari ez zaio behar bezalako arretarik jarri. Orain, berriz, morfologia konplexuko hizkuntzen tratamendu automatikoari ekitean, eta corpusetan oinarritutako analisi linguistikoak gora egin duen honetan, morfologia erabat beharrezkoa dela argi geratu da tamaina itzeleko hiztegiak saihesteko, lematizazioa egiteko, informazioa indexatzeko, informazioa bilatzeko…

---

[4] Aipatu behar da, dena den, zenbait sistema konputazionaletan ez dela hartzen morfema, baizik eta silaba, hitzaren osagai gisa (Cahill, 1990).

Gatozen orain euskarara edo edozein hizkuntza eranskaritara. *Mendi* adibidearekin jarraituz, mugatu/mugagabe marka erantsi behar diogu, artikulua alegia —mugatuetan—, eta horietako bakoitzari aldian aldiko kasua. Are gehiago, kasu bat baino gehiago har dezake genitiboak, esaterako, eta berriro mugatasuna eta kasua hartzeko prest egongo da. Horrela *ad infinitum* irits gaitezke teorian, praktikan birritan baino gehiago aplikatzea normala ez bada ere. Eta adjektiboei hiru gradu-marka erantsi behar diegu, horiek ere mugatasuna eta kasua har dezakete-eta, forma neutroaz gain.

Kasua	Mugagabea	Mug. sing.	Mug. plur.	Plur. hurbila
1) Absolutiboa	*mendi*	*mendia*	*mendiak*	*mendiok*
2) Partitiboa	*mendirik*			
3) Ergatiboa	*mendik*	*mendiak*	*mendiek*	*mendiok*
4) Datiboa	*mendiri*	*mendiari*	*mendiei*	*mendioi*
5) Inesiboa	*menditan*	*mendian*	*mendietan*	*mendiotan*
6) Leku-genitiboa	*menditako*	*mendiko*	*mendietako*	*mendiotako*
7) Adlatiboa	*menditara(t)*	*mendira(t)*	*mendietara(t)*	*mendiotara(t)*
8) Hurbiltze-adlatiboa	*menditarantz*	*mendirantz*	*mendietarantz*	*mendiotarantz*
9) Muga-adlatiboa	*menditaraino*	*mendiraino*	*mendietaraino*	*mendiotaraino*
10) Ablatiboa	*menditatik,* *menditarik*	*menditik*	*mendietatik,* *mendietarik*	*mendiotatik,* *mendiotarik*
11) Genitiboa	*mendiren*	*mendiaren*	*mendien*	*mendion*
12) Soziatiboa	*mendirekin*	*mendiarekin*	*mendiekin*	*mendiokin*
13) Instrumentala	*mendiz*	*mendiaz*	*mendiez*	*mendioz*
14) Motibatiboa	*mendi(ren)gatik*	*mendia(ren)gatik*	*mendiengatik,*	*mendiongatik,*
15) Destinatiboa **Gen.** **(gutxi erabilia)** **Ines.** **(gutxi erabilia)** **Adlat.** **(gutxi erabilia)**	*mendirentzat,* *menditako* *menditarako*	*mendiarentzat,* *mendiko* *mendirako*	*mendientzat,* *mendietako* *mendietarako*	*mendiontzat,* *mendiotako* *mendiotarako*
16) Banatzailea	*mendiko*			
17) Prolatiboa	*menditzat* *menditako*			

3. irudia. Izen arrunt bizigabeen euskal deklinabide-taula (Euskaltzaindia, 1993:518), [mugatu hurbila gurea da]

Zenbat sarrera zerrendatu beharko genuke euskaraz *mendi* bezalako izen arrunt bat hiztegian edukitzeko? Eta *txiki* gordetzeko? Muga gaitezen izenera: 3. irudiaren arabera, 17 kasu ditugu, baina 21 aldaera. Hauetako 17 aldaerek muga(gabe)tasun[5] osoa hartzen dute (17 x 4 = 68), adlatiboaren *–ra* eta *–rat* aldaera biak kontuan hartuta. Partitiboak, destinatiboko banatzaileak eta prolatiboak (azken honetako bi aldaerek) mugagabea bakarrik har dezakete (4 sarrera, beraz); ablatiboaren aldaerak mugagabea eta plurala (hurbila barne) har ditzake (3 sarrera) eta, azkenik, motibatiboaren mugagabeko eta mugatu singularreko aldaerak bi dira (*-ren* artizkia aukeran). Hortaz, 68 + 4 + 3 + 2 = 77 sarrera oinarrizko izango ditugu.

Baina, 77 sarrera hauetatik abiatuta, zazpi kasuk, lau muga(gabe)tasun-zutabeak kontuan hartuta, *-ko* har dezakete, eta 5 aldaerak ere bai (7 x 4 = 28; 28 + 5 = 33 forma berri). Horiek mugatu singularrean 17 forma berri osa dezakete, mugatu pluralean 19 eta plural hurbilean beste horrenbeste (17 + 19 + 19 = 55 forma gehiago). Aurreko 33ei lotuta, beraz, 88 forma ditugu.

[5] Alegia, mugagabea, mugatu singularra, plurala eta plural hurbila.

Eta jabego- eta leku-genitiboek mugatasuna (3 zutabe, singularra, plurala eta plural hurbila) har dezakete (2 x 55 ((17 x 3) + 2 + 2) = 110).

Hortaz, bigarren maila honetan 77 + 88 + 110 = 275 sarrera izango genituzke. Eta *mendikoan* modukoak aztertuko genituzke, testu-hitz arruntak, nolanahi ere.

Izen arrunt bizigabeekin egin dugu hau, baina bizidunekin beste horrenbeste beharko genuke, bereziekin beste modu batera; adjektiboei horrezaz guztiaz gain gradua erantsi beharko genieke, eta hortik berriro deklinabide-sistema martxan jarri; zenbatzaile zehaztuei ordinalak eta banatzaileak banaka gehitu, adizki jokatuak menderagailu eta guzti zerrendaratu, etab. Baina, guztia biltegiratzeko lekurik izango al genuke? Eta, izatekotan ere, nolako analisi-abiadura lortuko genuke? Zenbatgarren sarreran etsiko genuke? ... eta, sartuko al genituzke guztiak etsi aurretik?

Oinarrizko 77 formak bakarrik kontuan hartuta, Elhuyar-en *Hiztegi Entziklopedikoa*k (1993) dituen 50.000 sarrerak[6] sartzeko (izenak erreferentzia hartuta[7]), adibidez, 3.850.000 forma inguru beharko genituzke. Baina bigarren mailako 275 formetan oinarrituko bagina (oso normala, bestalde, hiztunarentzat maiztasun handikoak baitira horiek), 13.750.000 forma listatu beharko genituzke. Eraginkorra izango al litzateke hau?

Ondorioa argia da. Baina, ingelesera bueltatuz, ia kontrakoa erakutsi dugun arren, hark ere badu analizatzaile morfologikoaren beharra: flexio-morfologia mugatua du, bai, baina eratorpena oso aberatsa eta konplexua. Adibidez: *compute* sarreratik erator daitezke, besteak beste, *computer, computerize, computerization, recomputerize, noncomputerized*. Eratorri guztiak lexikoan zerrendatzea ezinezkoa da. Euskaraz ere antzeko zerbait dugu. Are gehiago, anbiguotasuna handia da ingelesez ere, aditza eta izena, esaterako, homografoak baitira gehienetan (*book, (to) book*). Lexikoan bakoitzak bere informazio egokia izango du, baina sintaxira pasa aurretik desanbiguatu beharra dago baliagarri izango bada.

Uste dugu honekin nahiko frogatuta geratu dela morfologiaren tratamendu automatikoaren beharra.

Morfologiaren azterketarekin batera datorren beste ikergaia *etiketatzea* da. Badago, besterik gabe, morfema zatiak eta euren arteko lotura erregulatzen duten erregelak baino ez zehaztea, alegia, hitzaren osaera morfologikoa hitz zatika bakarrik errepresentatzea. Eta, bestetik, badago morfemei eduki linguistikoa ematea, morfema bakoitza kategoria linguistikoen arabera deskribatuz. Hain zuzen horrelakoak dira egungo sistema gehienak. Horietarako, orduan, beharrezkoak dira kategoria edo *etiketa-sistemak.* Etiketa batzuk berdinak dira hizkuntza guztietarako, baina beste zenbait, hizkuntza jakinari lotuago egon ohi dira, berari dagozkion ezaugarri espezifikoek eraginda.

Horren ondorioz, hitz berak osaera morfologiko modu bat baino gehiago izan dezake. Errepara diezaiogun, esaterako, *dantzari* hitzari: batetik, ‘dantza (iz.) + -i (datiboa)’ gisa azter genezake, baina bestetik ‘dantzari (iz.)’ edo ‘dantzari (iz.) + ø (absolutibo mugagabe)’ gisa. Ordea, hitz hori agertzen den

[6]Beste 7.500 azpisarrerak alde batera utzita.

[7]Izenak hartu ditugu erreferentziatzat, kopuru aldetik kategoria honetakoak direlako ugarienak eta adjektiboen kopuru itzela adberbioen analisi kopuru murritzarekin oreka daitekeelako. Adibiderako besterik ez dira datuok, sor daitezkeen kopuru adierazgarriez ohartzeko bakarrik.

testu zatian, bata edo bestea izango da, eta hortaz, horixe bera hautatu beharra dago. Honenbestez, hitz baten analisi posible guztien berri ematen dugun bezalaxe, analisi horietatik guztietatik testuinguru jakin horri dagokiona zein den esan behar dugu. Horixe da, hain zuzen ere, etiketatzaileen zeregina, eta horrexegatik kokatzen da etiketatze-prozesu hau morfologiaren eta sintaxiaren artean.

Segidako puntuan, sarrera honetan laburki aipatu ditugun oinarrizko kontzeptuak sakonkiago aztertuko ditugu. Ondoren, morfologia konputazionalki tratatzeko dauden zenbait hurbilpen ikusiko ditugu, eta azkenik, gaur egun morfologiaren tratamenduak duen egoera eta eraiki diren zenbait sistema erakutsiko ditugu.

2.2 Oinarrizko kontzeptuak

2.2.1 Morfologiaren konplexutasun mailak

Abia gaitezen euskara, ingelesa eta gaztelaniako adibide banarekin konplexutasun mailez egokiago jabetzeko:

	euskara	**ingelesa**	**gaztelania**
Adibidea	lagunei	(to the) friends	(a los) amigos

Adibideotan ikus daitekeenez, euskaraz "–ei"[8] atzizki flexiboak mugatasun, numero eta kasuaren berri ematen digu; ingelesez, pluralaren adierazle den *-s* atzizkia dugu; eta azkenik, gaztelaniaz, genero maskulinoaren adierazle den *-o* atzizkia eta numeroaren adierazle den *-s* atzizkia ditugu. Alegia, euskaraz, lemaz gain beste hiru informazio mota ageri zaizkigu; ingelesez bakarra, eta gaztelaniaz bi.

Ikus dezagun adibide are konplexuago bat:

	euskara	**ingelesa**	**gaztelania**
Adibidea	lagunekikoa	(the one related to) friends	(el respectivo a los) amigos

Ikus daitekeenez, ingelesez eta gaztelaniaz lehenengo adibidean genuen informazio bera aurkitzen dugu tratatzen ari garen hitzean (*friends*, *amigos*). Euskaraz, berriz, lau kasu ageri zaizkigu segidan: edutezko genitiboa (*-en*), soziatiboa (*-kin*), leku-genitiboa (*-ko*) eta absolutiboa singularrean (*-a*); baina, gainera, jakin badakigu, azken bi kasuen artean elementu ezkutu bat (elipsia) ere badagoela, aldez aurretik aipatu

[8] Morfema honen azpian, 'ak + i' dagoela badakigun arren, azalpenen erraztasunerako,–e– plural markatzat joko dugu zuzenean.

den zerbaiti erreferentzi egiten diona. Beste bi hizkuntzetan, berriz, informazio hori aparteko hitzetan ageri diren morfemek ematen dute.

Lema bakoitzak ondoren har ditzakeen atzizki guztiekin osatutako hitzak osoki sartu behar balira hiztegian, hiru hizkuntzotan oso tamaina desberdineko hiztegiak izango genituzke; euskaraz, aipatu dugun bezala, askozaz ere handiagoa litzateke. Eta horrek eraginkortasunari begira arazo handiak ekarriko lituzke. Horrexegatik ez da zalantzan jartzen hitzak morfemetan segmentatu eta lotzeko sistemen beharra.

2.2.2 Hitzen osaera morfologikoa: flexio-morfologia, eratorpen-morfologia eta elkarketa

Esan dugun bezala, morfema guztiak ez dira mota berekoak.

Flexio-morfologian, lema batek eta flexio-marka batek hartzen dute parte. Hori horrela izanik, flexio-morfologia sintaxiak eskatzen du eta sintaxiaren ezaugarrien berri ematen du (numeroa, mugatasuna, funtzioa); hitzak sintaxian txertatuta azaltzen direla san daiteke. Horretaz gain, normalean erregularra da lotzen zaion kategoriaren arabera; hau da, izen guztiek, edota mota orokor jakin bateko izen guztiek, flexio-forma bera onar dezakete, eta orobat adjektibo, aditz, eta gainerako kategoria lexiko guztiekin. Eta, azkenik, oinarrizko lemaren kategoria ez dute aldatzen; esan bezala, funtzioa baino ez da zehazten. Adib.:

lagunei (eusk.), *(to the) friends* (ingl.), (a los) *amigos* (gazt.)

Ikus dezakegunez, hiru hizkuntzotan hitzaren kategoriak izena izaten jarraitzen du, eta jakin badakigunez, flexio jakin hau izen guztiek har dezakete.

Eratorpen-morfologian, berriz, lema batek eta eratorpen-atzizki batek hartzen dute parte. Eratorpen-atzizki horrek, oinarriko lemari lotuta, hitz berri bat sortzen du, askotan kategoria-aldaketa bat gertatzen delarik, baina flexio-atzizkia ez bezala, ezin zaio edozein oinarriri lotu. Eratorpen-atzizkien zeregina hitz berriak sortzea izaki, ez du sintaxi-kontuen berri ematen. Adib.:

erosle (eusk.), *buyer* (ingl.), *comprador* (gazt.)

Hiru hizkuntzetan, eratorpen-atzizkia (*-le* euskaraz, *-er* ingelesez eta *–dor* gaztelaniaz) aditzari lotzen zaio, eta lotura horren ostean, izen kategoriako hitz berri bat sortzen da. Ingelesez eta gaztelaniaz, ordea, atzizkiak muga gehiago jartzen dizkio aditzari: aditzak erabilera iragankorra behar du izan, oro har: *comer (ingl.), *venidor (gazt.); euskaraz, berriz, zabalagoa dirudi, *etorle*, *joale*, *egoile*… zilegi baitira, nahiz eta seguru asko emankortasunari begira mota horietakoek jada ez lekurik eduki.

Elkarketan, azkenik, bi lemak edo gehiagok hartzen dute parte. Bi lema hauen elkarketak, eratorpenean bezalaxe, hitz berri bat sortzen du, eta hartan bezalaxe, ez du sintaxiaren ezaugarrien berri ematen. Horretaz gain, ezin da edozein lema jarri konposizioan, eta hitz berriaren kategoria konposizio motaren araberakoa da. Adib.:

hanka-motz (eusk.), *gammy-legged* (ingl.), *paticojo* (gazt.)

Ikus dezakegunez, euskaraz eta gaztelaniaz izen + adjektibo segida dugu; ingelesez, berriz, adjektibo + izen segida, eta gainera izenari partizipio-marka eransten zaio. Hiru hizkuntzetan, ordea, kategoria bereko hitzak sortzen dira (adjektiboak).

Flexio-morfologia, eratorpen-morfologia eta elkarketa, gainera, hitz berean agertzen dira askotan. Hortxe ditugu *erreferentziakidetasunaren*, *berrerabilgarritasun-neurria*, *bidelagun-taldea*… hitzak, esaterako. Horrek guztiak hitzaren osaera morfologikoaren zailtasuna eta beharra azpimarratzen ditu.

Bestetik esan behar da, hirurak ondo bereizita eman baditugu ere, badirela kasuak hiruen arteko mugak gainditzen dituztenak, eta zatiketa hau zirriborratzen dutenak. Euskaraz, esaterako, badira orain arte eratorpen-atzizkitzat hartu izan diren atzizkiak –esaterako, *-dun*–, erakusten dituzten erregulartasunak direla-eta, atzizki flexiboen artean lekutzera eramaten dutenak. Badira ere konposizio eta eratorpen artean dauden morfema-hizkiak, hala nola, *alde, arte, azpi, behe, kide, gai, gune, orde*… Gauzak, hortaz, ez dira guztiz-guztiz argiak.

2.2.3 Morfemen arteko lotura: morfotaktika eta morfofonologia

Morfema guztiak mota berekoak ez izateaz gain, morfemen arteko lotura ere ez da beti berdin egiten. Beraz, hitzen osaera morfologikoan kontuan izan beharreko beste atal nagusia morfema hauen arteko kateatzea da. Kateatzea bi zentzutan:

- *Morfotaktika.* Ikusi dugu, morfologia desberdinen arabera, morfema-segiden murriztapenak desberdinak direla. Murriztapen horien gauzatzea *Morfotaktika* delakoaren bidez egiten da. Hau da, Morfotaktikak, murriztapenak ezartzen ditu, jokabide egokia bakarrik posible eginez. Arestiko puntuan erabilitako adibideei helduaz, *zuhaitz* hitzari *-ei* datibo plurala jarraitzea zuzena eta beharrezkoa dela adierazi behar da; ez, ordea, *-tasun* atzizkia, esaterako. Honenbestez, oso ondo eta sistematikoki jakin behar da zein morfemak zein morfemari jarrai diezaiokeen. Eta, noski, horretarako lehenik eta behin jakin beharra dago zerk osatzen duen morfema bat eta zerk ez.
- *Morfofonologia.* Morfemak ez dira beti zuzen-zuzenean bere horretan lotzen. *Zuhaitzetik* esaten dugunean, badakigu azpian ‘zuhaitz + tik’ morfema-segida dugula, baina hizkuntzaren baitako joera fonologikoak direla-eta, bi morfema horien artean, -e- epentetiko bat sartu beharra dago. Morfotaktika horretan, hortaz, aldaketa hauek ere jaso behar dira. Aipatutako adibide hau, kateatze nahikoa sinplea da, eta horrelakoak dira, oro har, gure inguruko hizkuntzetan aurkitzen ditugunak. Baina horien aldean, badaude hizkuntza batzuetako –arabiera, esaterako– *erro-patroi* eredu konplexuagoak; hau da, atzizki bat lotzean hitzaren erroa bera ere aldarazten dutenak. Edota bokal-harmonia erakusten duten hizkuntzak, non puntu batean gertatzen den bokal baten aldaketak ondoko bokalen aldaketa ere eragin dezakeen. Morfemen arteko loturetan, honenbestez, fenomeno desberdinak gerta daitezke, arrazoi desberdinengatik.

2.2.4 Etiketatzea

Esan dugun bezala, morfologiari ekiteko, zenbait bide daude: hizkien bereizketa hutsa, batetik, eta hizkiak kategoria linguistikoetan bereiztea, bestetik. Teoria linguistikoak eraginda, baita hizkuntzaren egiturak

nahiz morfologia ondorengo sistemak eraikitzeko konputazio-ikuspegiak ere, bigarren eredua bihurtu dute erabiliena. Erabiltzen den kategorizazioaren arabera, eta hizkuntzaren izaeraren arabera, hitz baten osaera morfologikoa sinpleagoa edo konplexuagoa izango da[9]. Euskarazko adibide batera joaz:

> "*handiena*" hitzaren morfemakako osaera hutsa hauxe litzateke:
>
> handi+en+a.
>
> Hitz beraren kategorizazio guztiak kontuan hartutako analisia, berriz, beste hau:
>
> handi[KAT_ADJ]+en[KAT_GRA][GRM_SUP]+ a [NUM_S][MUG_M]
> handi[KAT_ADJ]+en[KAT_DEK] [KAS_GEN] [NUM_P][MUG_M] + a [NUM_S][MUG_M]
>
> Alegia, *handi* lema bera izanda ere, *-en* morfema, *graduatzaile superlatiboa* edo *genitibo mugatu plurala* izan daiteke.

Morfologiaz harantzagoko analisietara jotzeko, bigarren eredua da gailentzen dena, gerorako (sintaxia, semantika…) beharrezkoa izango den informazioa zehazten baitu.

Hizkuntza-unitate horiek kategorizatzeko, orduan, kategoria- edo etiketa-sistemak eraikitzen dira, eta unitate morfologikoak, hain zuzen, horren arabera definitzen dira lexikoan. Etiketa asko hizkuntza guztietan topatzen baditugu ere, zenbaitzuk hizkuntzen ezaugarriei lotuago daude: euskarak, hots, hizkuntza morfologikoki konplexuek, esaterako, kategoria ez-aske asko dituzte; sinpleagoek, berriz, askozaz ere gutxiago.

Horrek baina, analisien aberastasunarekin batera, anbiguotasuna dakar ondorio gisa, *handiena* adibidean ikusi dugun bezala, hitz-forma berak analisi morfologiko bat baino gehiago izan baititzake.

Hori dela eta, desanbiguazio-metodoak aplikatu behar dira, aplikazio desberdinetarako funtsezkoa izango baita hitz horri dagokion analisi morfologiko egokia esleitzea: dagokion kategoria eta informazio morfosintaktikoaz ematea.

Hala eta guztiz ere, zein den helburua, desanbiguazioa ere maila desberdinean egin daiteke. Hala, helburua lematizazioa, indexazioa edo testuetan bilaketak egitea bada, nahikoa izango da kategoria mailan egitea. Helburua sintaxia bada, ezaugarri morfosintaktikoen desanbiguazioa egin beharko da. Helburua semantikoa bada, kategorien azpikategoriei arreta berezia jarri beharko zaie.

Desanbiguazio-prozesu honetarako formalismo erabilienak ezagutza linguistikoan oinarritutako egoera finituko mekanismoak dira, testuinguruari loturikoak. Formalismorik ezagunenetariko bat, eta euskararentzat ere baliatu dena, Murriztapen Gramatika delakoa dugu (MG) (Karlsson, 1990). Formalismo horrek testuinguruan oinarritutako erregelak definituz hitz-formen desanbiguazioa bideratzen du, horretarako ezagutza linguistikoaz baliatzen delarik. Esan behar dugu formalismo horiek, desanbiguazio-mekanismoez gain, sintaxira hurbiltzeko pausoak ere eskaintzen dituztela, esaldiko osagaiei funtzio sintaktikoa esleitzeko ahalmena baitute. Horren bidez, perpausa sintaktikoki etiketatuta

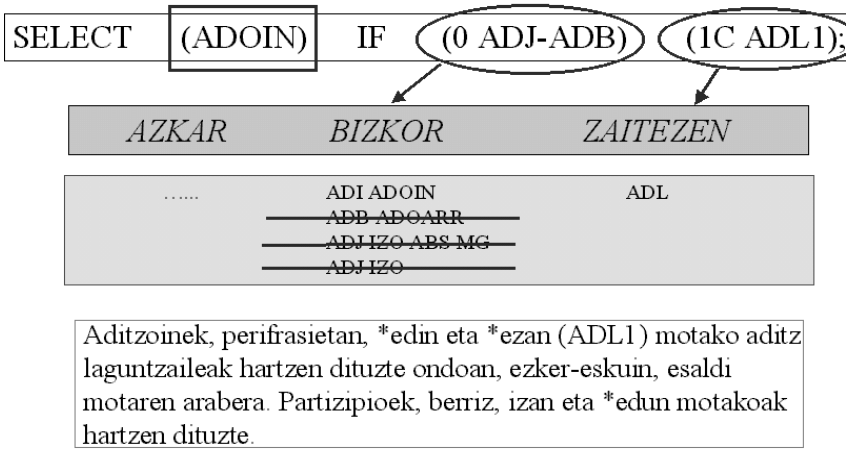
---

[9] Esaterako, IXA taldean, morfologian eta sintaxian erabiltzen diren etiketa guztiak, Aldezabal 2004 barne-txostenean ikus daitezke (http://ixa.si.ehu.es/Ixa/Argitalpenak/Barne_txostenak).

geratzen da. Era berean, funtzio sintaktikoen etiketetatik abiatuz, posible da perpauseko sintagmak identifikatzea.

Etiketatzaileek, honenbestez, hitzen interpretazio posible guztien artean desanbiguazioa egiten dute, baina horretaz gain, etiketatze sintaktikoa ere bideratzen dute, eta ondorioz, analizatzaile sintaktiko partzial bihurtzen dira.

Ikus, adibide gisa, euskarazko MGn desanbiguaziorako baliatzen den erregela bat:



4. irudia. Murriztapen Gramatika bidezko desanbiguatze-erregela bat

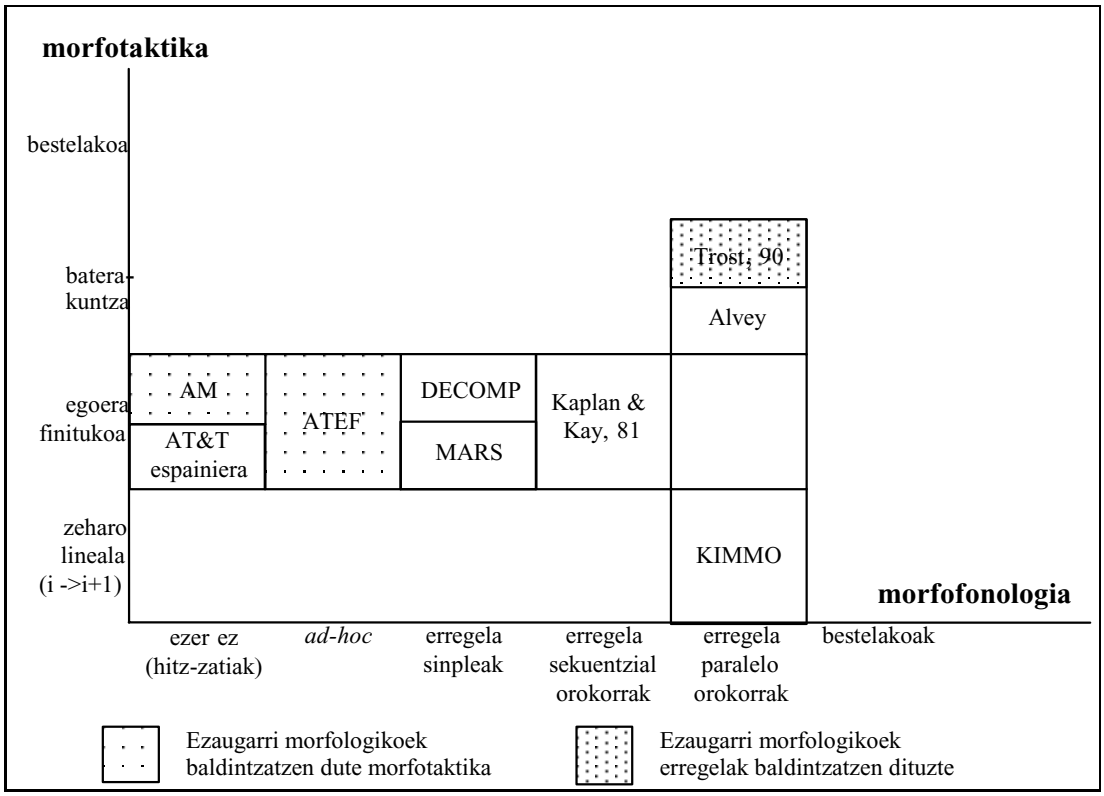
Horren antzeko erregelen bidez, funtzioak esleitu eta horien gainean, sintagmak osatzeko etiketak ezartzen dira.

Etiketatzaileak, hortaz, funtsezkoak dira lematizaziorako, baina baita sintaxian aurrera egiteko. Halere, esan behar da, lematizazioari gagozkiola, era bateko lematizazioa egiteko badirela desanbiguazioaz harantzagoko beharrak, hala nola, hitz anitzeko unitate lexikalen identifikazioa (lokuzioak, pertsona-izen osoak etab.).

2.3 Hurbilpenak

Morfologia aztertzeko hainbat hurbilpen daude. Guk, hemen, eraiki diren zenbait sistematan erabilitakoak azalduko ditugu laburki (xehetasunerako jo bedi Alegria & Urkia (2002) liburura).

5. irudian ikus daitezke hurbilpenok.



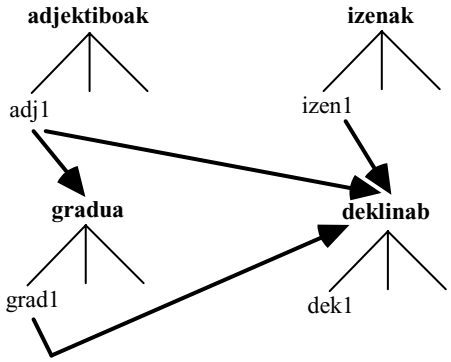
5. irudia. Prozesadore morfologikoen sailkapena

Oro har, hiru hurbilpen bereizten dira: *egoera finitukoa, baterakuntza-mekanismoetan oinarritutakoa* eta bestelakoak.

*Egoera finitukoa*n, morfemen arteko erlazioak grafo-eran ikus daitezke, nodoak, morfemak eta arkuak onartutako kateatzeak izanik.

Adibidez, eta sinplifikazio bat eginez, euskarazko adjektiboek eta izen arruntek deklinabide-atzizki berberak hartuko dituzte, baina lehenek baino ezin dute gradu-flexiorik hartu. 6. irudian ikus daiteke lexiko sinplifikatu honen eraketa.

LEXIKOA **izenak**	JARRAITZE-KLASEA **izen_jk**
ize1 izen_jk ...	deklinab
ize2 izen_jk ...	
...	JARRAITZE-KLASEA **adj_jk**
LEXIKOA **adjektiboak**	deklinab
adj1 adj_jk ...	gradua
adj2 adj_jk ...	
...	
LEXIKOA **gradua**	
grad1 izen_jk ...	
grad2 izen_jk ...	
...	
LEXIKOA **deklinab**	
dek1 #[10] ...	
dek2 # ...	
...	



6. irudia. Lexiko-sistemaren erazagutzearen eta egituraren adibidea

*Baterakuntza-mekanismo*ek sintaxian erabili ohi diren ezaugarrietan oinarritutako gramatikak aintzakotzat hartzen dituzte, eta ondorioz, malguagoak dira. Tratamendu morfologiko —edo morfosintaktikoa— errazten dute, baina konplexuagoak dira konputazioaren ikuspuntutik. Morfemek osatzen duten hitzaren informazioa herentzian oinarritutako printzipioen bidez lortu ohi dira. Mekanismo hauek erabiltzeko, jakina, morfemek linguistikoki kategorizatuta behar dute egon.

Adibidea: Iz. arr. / adj. + kasu-marka: etxe + a

Adib.: etxea

	etxe	a
EREM:BAL	kat: IZE	kas: ABS
	azpik: ARR	num: S
	biz: -	mug: M
	zenb: +	fs1: @SUBJ
	pl: NULL	fs2: @OBJ

⇓

Baterakuntza-ekuazioak

[10] # sinboloak jarraitze-klase hutsa adierazten du.

⇓

etxea

EREM:BAL

kat: IZE
azpik: ARR
biz: -
zenb: +
kas: ABS
num: S
mug: M
fs1: @SUBJ
fs2: @OBJ

Morfemen arteko aldaketa morfofonologikoen berri emateko, berriz, bibliografian bi metodo gailentzen dira: bata, orain dela urte batzuk ohikoak ziren programa bidezko metodo *ad hoc*ak, eta, bestea, gaur egun arrakastatsu bihurtu den *bi mailako* formalismoa.

Metodo *ad hoc*ek, kasuan kasuko morfemei dagozkien aldaketak zerrenda huts gisa definitzen dituzte hiztegian, morfema beraren alomorfoak bi sarrera gisa definituz.

Esaterako: *egiten, egingo, etortzen* eta *etorriko* formak lortzeko, ondoko lemak eta morfemak izan beharko genituzke landuta lexikoan: egi, egin, etor, etorri, ten, tzen, go, ko.

Bi mailako formalismoa, berriz, bi mailatan mugitzen da: azalekoan eta lexikoan. Eta horrek, hain zuzen, ahalbidetzen du morfema beraren gauzapen fonetikoak sarrera bakarrari lotuta agertzea. Formalismo honek bi osagai ditu: sistema lexikoa eta erregelak. Hala, erregelek, morfemen sakoneko mailan definitzen diren diakritikoetan eta morfemen testuinguruan oinarrituta, azaleko gauzapena bideratzen dute.

Esaterako *ber*- aurrizkiaren “r”a gogorra da; hots, *ber*-i lema bat lotzen zaionean, “rr” ageri zaigu azalean (ez “r”); horretaz gainera, lema “h”z hasten bada, “h” hori galdu egiten da; eta azkenik, lema “h” ez den beste kontsonante batez hasten denean, “e” “i” bihurtzen da. Hori guztia ondorengo bi erregelen bidez konpontzen dugu lexikoan lema bakarra izanda (erregelak sinplifikatuta daude; jo bedi Alegria & Urkia 2002: 39ra, azalpen zehatzagorako):

h:0 <=> #: b e R (0:r) +: _ ;

lexikoa: b e R + h a s i
| | | | | | | | |
azala: b e r r ø ø a s i

e:i <=> #: b _ R +: Konts ;

lexikoa: b e R + f i n
| | | | | | |
azala: b i r ø f i n

2.4 Egungo egoera

Zenbait hizkuntzatan egin diren analizatzaile morfologikoak Morfotaktikan gertatzen diren morfema-kateatze eta aldaketa morfofonologiko horiek bideratzeko moduaren, eta tratatzen dituzten morfologia moten arabera bereizten dira.

Ondoren zerrendatzen ditugunak ordena kronologikoan ulertu behar dira, eta adibide adierazgarri batzuk baino ez dira.

- DECOMP (Allen eta beste, 1987) (ingeleserako egina).
- ATEF, Grenobleko GETA taldeak (GETA, 1982) egindakoa (hizkuntza askotan aplikatu dena: frantsesa, errusiera, alemana eta hizkuntza asiatikoak…).
- Kay-ren *chart parsing*a (Kay, 1977), garai batean estandartzat hartu zena.
- KEÇI (Hankamer, 1986), turkierarako prestatutako egoera finituko analizatzailea.
- Samba.
- Bi Mailako Morfologia (Koskenniemi, 1983) edozein hizkuntzari aplika dakiokeen formalismoa, suomierarako gauzatu arren, berehala izan zuena ingeleserako bertsioa (KIMMO) (Karttunen-ek (1983) eginda) eta euskara lantzeko hautatu dena.
- AMPLE (Weber *et al.,* 1988), esplorazio morfologikorako prestatua; Tzoukermann eta Liberman-ek (1990) proposatutako eredu sinple baina eragingorra.
- MOLUSC (Cahill, 1989, 1990), silaban oinarritutako sistema.

Alegria & Urkiaren lanean (2002) horiek guztiak zehazkiago aztertzen dira. Guk hemen euskararen morfologia deskribatzeko erabili den Bi Mailako Morfologiaren zenbait zertzelada emango ditugu.

Koskenniemi-k sortutako eredu honek gramatika sortzailean proposatzen diren azaleko eta sakoneko egituraren pareko diren *lexiko* eta *azaleko* mailak bereizten ditu. Bada, halere, desberdintasunik: gramatika sortzailean transformazio bidez pasatzen da egitura batetik bestera, baina bi mailako ereduan ez dago tarteko egoerarik. Fonologia sortzaileko berridazketa-erregelen ordez, erregela paraleloak proposatzen ditu Koskenniemi-k. Hala, bere sisteman, bi osagai dira oinarri: sistema lexikoa eta erregelak.

Sistema lexikoak morfema multzoa definitzen du, morfemen artean egon daitezkeen kateamenduen arabera sailkapena eginez. Azpilexikoen multzoa eta erroen eta hizkien sekuentzia posibleak erregulatzen dituzten jarraitze-klaseek osatzen dute sistema hau. Azpilexikoek ezaugarri berdineko elementu lexikoak (lemak, aurrizkiak, atzizkiak…) biltzeko balio dute, eta egitura bera dute gainera: identifikatzen dituzten izena eta sarrera multzoa. Sarrera bakoitzak, bere aldetik, hiru eremu ditu:

- Adierazpen lexikoa: karaktere-sekuentzia bat da. Karaktere hauek azaleko karaktereak edo hautapen-markak –alegia, diakritikoak, maila lexikoan bakarrik mugitzen direnak– izan daitezke. Azkeneko horiei azaleko beste karaktere batzuk egoki lekizkieke erregelen bitartez.

- Dagokion jarraitze-klasea: zenbait azpilexikoi edota beste jarraitze-klase batzuk biltzen dituen identifikadorea da hau. Jarraitze-klasean biltzen diren osagaiak dira definitutako sarreraren atzetik ager daitezkeen bakarrak.
- Sarrerari dagokion *informazio morfologikoa.*

Beraz, jarraitze-klaseak hitz batean ager daitezkeen morfemen arteko konbinazio posibleak definitzeko mekanismoen oinarri dira.

Erregelek hiru osagai dituzte:

- *Korrespondentzia,* edo karaktere-bikote bat, lehenengoa lexiko mailakoa eta bigarrena azaleko mailari dagokiona (adib. i:j)
- *Testuingurua,* korrespondentzia gertatzen den kasuak mugatzen dituena, aurreko eta ondorengo karaktereen arabera (lc_rc). Lc ezkerreko testuinguari dagokio (left context), '_' lantzen ari garen karakterea da eta rc (*right context*) karaktere horren ondoren, hau da, eskuinera datorren testuingurua.
- *Eragilea,* testuinguruaren eta korrespondentzian adierazitako bikotearen artean dagoen erlazio mota finkatzen duena. Erlazio hau lau modutara murritz dezake: testuingurua mugatuz (⇒ ikurraz adierazita), azalekoa derrigortu dezake (⇐ ikurraz adierazita); bi horiek batera (⇔ ikurraz adierazita); eta, azkenik, debekua ezar dezake (/⇐).

Adibidez:

i:j	⇒	b:b	–	e:e
korrespondentzia	eragilea	ezkerreko testuingurua		eskuineko testuingurua

Horrek ondokoa adierazten du: lexikoko “i”, azaleko “j” bilakatuko da baldin eta ezkerrean “b” eta eskuinean “e” badu.

Erregela hauen beste ezaugarri nagusi bat da ez dagoela ordenarik euren artean, eta ondorioz, paraleloan –alegia, aldi berean– aplika daitezke.

Morfologia-eredu honi jarraituz, IXA taldean garatu dugun lexikoiak (Euskararen Datu-Base Lexikala (EDBL)) 83.070 sarrera ditu (lema eta morfema), 201 azpilexikoi eta 160 JK. Erregelak, berriz, 23 dira[11].

[11] Datu hauek 2004ko uztailaren 27an ateratakoak dira.

3 Sintaxia

3.1 Sarrera

Sintaxian, osagai sintagmatikoen arteko loturak aztertzen dira; alegia, hitzaren muga gaindituz, hitz horien konbinazioek sortzen dituzten unitate handiagoak dira unitate aztergaia. Unitate sintagmatiko horiek sortzeko mekanismo sintaktikoak eta sortutako unitate horiek perpausean harremanetan jartzeko duten jokamoldea dira axola duena. Zuriunetik zuriunerako karaktere-segida, hortaz, ez da jadanik muga, eta perpausaren eremuetan murgiltzera pasatzen gara. Morfologia aztertzean ikusi dugun bezala, ordea, hitz batek informazio ugari dakar euskara bezalako hizkuntza eranskarietan, eta horrek hitz mailan sintaxi-fenomenoak ere aztertzea posible egiten du. Hortxetik dator morfosintaxia diziplina, morfologia eta sintaxia oso loturik daudela erakusten duena.

Unitate sintagmatiko hauen harremanak eta perpausa osatzeko mekanismoak, ordea, ez daude guztiz formalizatuta teoria mailan. Horrela da gehien ikertu diren hizkuntzetan eta, are gehiago, euskara bezalakoetan. Linguistika teorikotik testu errealen analisira pasatzeko mementoan egoera zailagoa da, testu horietan linguistikoki deskribatu gabeko hainbeste egitura mota agertzen baitira. Hori dela eta, zenbait egile (Sampson, 1987) ohiko gramatiken baliagarritasuna zalantzan jartzera iritsi dira. Puntu horretara joan gabe eta hizkuntza baten sintaxi osoaren deskribapena oraindik burutu gabeko lana izanik ere, sintaxiaren tratamendu automatikoan aurrera egin da sintaxi partziala deritzona bideratuz gehienbat.

Bestalde, sintaxiaren deskribapen partziala izanda ere, gramatika sintaktiko batek lortutako egitura guztiak aplikatuz gero, edozein esaldi arruntentzat aukera asko sortzen dira, gehienak zentzugabeak testuingurua aztertuz gero. Beraz, esaldi mailako tratamenduan, morfologian ikusi dugun bezalaxe, beste ikergai nagusia anbiguotasuna ezabatzea da. Helburu nagusia esaldi bakoitzeko analisi bakarra lortzea izango da.

Sintaxiaren tratamendurako hurbilpenek, hortaz, bi arazo gainditu beharko dituzte: esaldi osoaren interpretazio posible guztiak identifikatzea eta beraien artean aukeratzea. Horiek biak bi modutara ebatz daitezke: ezagutza linguistikoa eskuz kodetuz edo automatikoki lortutako ezagutza erabiliz. Hortik, sintaxiaren prozesamenduan, hiru hurbilpen nagusi bereizten dira: eskuz kodetutako ezagutza linguistikoan oinarritutako hurbilpenak, estatistikan edo metodo automatikoetan oinarritutako lanak, eta bien konbinazioak.

Ondorengo puntuetan, alde batetik, sintaxi konputazionalari ekiterakoan kontuan izan behar diren hainbat alderdi azalduko ditugu, horretarako sintaxi teorikoarekiko aldeak azpimarratuz; eta bestetik, sintaxi konputazionala lantzeko dauden hiru hurbilpenen berri emango dugu, horiek baliatzen dituzten hainbat formalismo ere aurkeztuz.

3.2 Oinarrizko kontzeptuak

Ezagutza linguistikoa kodetzeko, *gramatikak* erabili ohi dira. Horiek, sarreran aipatutako bi problemak (analisiak lortu eta zuzena aukeratu) ebazteko erregelen edo prozeduren deskribapenak dira, gehienetan gizakiak eginak. Gramatika bat idazteko orduan, hainbat irizpide har daitezke, horietako batzuk elkarren kontrakoak eta besteak, aldiz, osagarriak direnak. Irizpideen jatorria, normalean, sintaxira hurbiltzeko hautatzen den ikuspegiaren araberakoa da. Batzuk teoria linguistikoak egiaztatzeari erreparatuta eginak dira batez ere, eta beste batzuk ikuspegi praktikotik baliagarri izateari erreparatuta gehiago. Horrek ez du esan nahi analisi linguistikoak ez direla erabilgarriak LNPn, baizik eta teoria linguistikoen (eta hauetan oinarrituriko formalismoen) eta aplikazioen artean lotura bat definitu behar dela.

Ikus ditzagun ondoren bi ikuspegi horien arteko desberdintasun nagusiak, eta hortik sortzen diren zenbait oinarrizko kontzeptu.

3.2.1 Unitate aztergaia

Aipatutako bi ikuspegien unitate aztergaia esaldia[12] bada ere, esaldi horren izaera desberdina da hainbat alderdiri erreparatuta:

	Hizkuntzaren teoria egiaztatzea helburu	**Aplikazio errealetarako baliagarria izatea helburu**
Esaldi kopurua	datu gutxi, “laborategikoak”	datu asko
Esaldien jatorria	esaldi asmatuak	esaldi errealak
Esaldien zuzentasuna	esaldi gramatikalak	esaldi gramatikalak nahiz ez-gramatikalak
Testuingurua	ez da tratatzen	tratatu behar da

Aplikazio errealetan baliatzeko hurbilpenek eguneroko jardunean esku artean darabiltzagun testu-masak edo corpusak dituzte abiapuntu. Hori dela eta, datu asko maneiatu behar ditu, esaldiak benetakoak (errealak) dira, ekoizpen zuzen nahiz okerrak daude, eta dagokien testuinguruan ulertu/analizatu behar dira. Hizkuntzaren teoria egiaztatzea helburu dutenak, berriz, datu gutxi batzuk eta normalean norberak asmatuak tratatzen ditu, ondorioz ez ditu esaldi okerrak aztertu behar, eta azkenik, ez dira inongo testuingurutan ulertu behar, sortzez duen interpretazioan baizik.

Corpuseko adibide batzuk:

> Kurban **agertu** zinenean lagun guztiok geunden. (EusCor[13])
>
> Beste ehun milioi urte igaro ziren, aldi karboniferoa ailegatu zen eta Lurrean baso trinkoak **agertu** ziren, zeinetan iratzeak, azeri-buztana eta likopodioa hazten bait ziren. (EusCor)

[12] *Esaldi*tzat jotzen da puntutik punturako testu zatia.

[13] EusCor laburdura erabili dugu *XX. mendeko euskararen corpus estatistikoa* adierazteko.

> Zer edo zer idaztera zihoan baina haren luma lehen letra idaztera zihoanean, zergatik jakiteke, dardar jarri zitzaion eskua, ezin izan zuen ezer idatzi, lehendabiziko letra hura arestian bizitako sentimendu haiek deformatzen zituen zerbait bezala **agertu** bailitzaion. (EusCor)

Bigarren eta hirugarren adibideek, esaterako, 24 eta 36 hitz dituzte hurrenez hurren, eta perpaus bat baino gehiagoz osatuta daude. Lehendabizikoak, bestalde, ez ditu hainbeste hitz, baina *Kurban **agertu** zinenean* perpaus zatian, aditz laguntzailearen bidez badakigu "zu" moduko subjektu bat suposatutzat eman behar dugula, testuan esplizituki agertzen ez den arren.

Aldiz, esaterako, GB teoriaren barruan sintagmen "hesiak" frogatzeko saioan, ondoko adibideak erabili ohi dira:

Nork esan duzu ikusi duela filma?

**Noren esan duzu filma ikusi duela Pellok?*

Ikus daitekeenez, adibide hauetan ez da inongo testuingururik, ez eta inongo elementu eliptikorik planteatzen; izatez, besterik gabe "esaldi osoa" esplizitu adierazten da, nola behar lukeen eta izango litzatekeen zehatuz. Hortaz, gramatika-arauak, besteak beste, hori definitzeari begira egingo dira. Ez dira arduratzen ikusitako corpuseko esaldien modukoetan dauden ezaugarriak errepresentatzen.

3.2.2 Sintaxi osoa vs sintaxi partziala

Izatez, eta sarreran aipatu dugun bezala, sintaxi osoaren deskribapena oraindik burutu gabeko lana da, eta horren ondorioz, sintaxi konputazionalean aurrera egiteko, *sintaxi partziala* deritzona landu da, hizkuntzaren teorian oinarritutako formalismoen *sintaxi osoa*ren parean. *Sintaxi partzial* terminoak teknika desberdinen multzoa definitzen du, eta analisi tradizionalaren informazioaren zati bat, ez guztia, lortzen du.

Har dezagun, esaterako, aipatu ditugun adibideetatik sinpleena.

> Kurban **agertu** zinenean lagun guztiok geunden

Esaldi honetan idealena litzateke dauden sintagma eta perpaus mailako erlazio guztiak adieraztea: lehenik "kurban agertu zinenean" denborazko mendeko esaldi bat dela jakin beharko genuke, eta, era berean esaldi hori "Kurban" eta "agertu zinenean" sintagmez osatua dela. Bestetik, "lagun guztiok" eta "geunden" ere sintagmak direla jakin beharko genuke; eta, azkenik, bi sintagma horiek mendeko denborazko perpausarekin batera esaldia osatzen dutela jakin beharko genuke:

> ((Kurban) (agertu zinenean)) (lagun guztiok) (geunden)

Hortaz, ordenagailuari horren guztiaren berri eman behar diogu gramatiken bidez. Esaldi horretan dauden sintagma eta erlazioen berri, alegia. Izan ere, hori gabe ordenagailuak ezin du jakin esaldi horretan agertzen diren sei hitzak zeri dauden lotuta. Informazio hori gabe, hasiera batean hitz guztiak guztiekin konbinatuz gerta daitezkeen egitura sintaktiko guztiak emango lituzke aukeran.

Adibide luzeagoetara jotzen dugunean, areagotu egiten dira zailtasunak, horrelakoetan gizakiak berak ere ez baitaki banaketa sintaktiko zuzena zein den. Har dezagun aipatu ditugun adibideetatik luzeena:

> Zer edo zer idaztera zihoan baina haren luma lehen letra idaztera zihoanean, zergatik jakiteke, dardar jarri zitzaion eskua, ezin izan zuen ezer idatzi, lehendabiziko letra hura arestian bizitako sentimendu haiek deformatzen zituen zerbait bezala **agertu** bailitzaion. (EusCor)

Nola banatu behar dugu esaldi hau? Ondoko modura, adibidez?

> (((Zer edo zer) (idaztera)) zihoan) baina ((haren luma) (lehen letra) (idaztera) (zihoanean)), (zergatik jakiteke), ((dardar) (jarri zitzaion) (eskua)), ((ezin izan zuen) (ezer) (idatzi)), (lehendabiziko letra hura) ((((arestian) (bizitako)) sentimendu haiek) (deformatzen zituen) (zerbait)) bezala) (**agertu** bailitzaion).

Bada, hori egitea zaila bada gizakiarentzat, are zailagoa da hori bera behar bezala deskribatzea ordenagailuarentzat, hitzak, sintagmak, perpausak, komak, puntuak... ondo definitu behar baitira testuinguru guztiak kontuan izanda.

Hori dela eta, analisia partzialki egiten da: sintagma zatiak osatzen saiatzen da, horien arteko erlazio batzuk markatuz, baina beste batzuk aske utziz. Hortik *partzial* (edo *azaleko*, alor konputazionalean) deitura.

3.2.3 Ikuspegi eraikitzailea vs ikuspegi murriztailea

Esan dugun bezala, analisi partzialean hainbat erlazioren berri ematen da. Zenbat erlazio aurreikusi, hainbat aukera izango ditugu zuhaitz sintaktikoak eraikitzeko. Hala, zenbait formalismok lexikoiko informazioa erabiliz esaldia analizatzeko aukera posible guztiak sortzen dituzte, eta gramatikariaren edota erregela sintaktikoen lana onartezinak baztertzea da. Horrelako formalismoak *murriztaileak* direla esaten da. Beste formalismo batzuk, ordea, hitzetatik abiatuta, beraien konbinazioaz bakarrik esaldiaren azken egitura eraikitzen saiatzen dira, era determinista alferrikako aukerak sortu gabe. Horiei formalismo *eraikitzaileak* esaten zaie. Hala ere, gramatikak eraikitzerakoan bi formalismoen konbinazioa egiten da, egitura sintaktiko guztiak ez daudelako hasieratik sortuta, eta gainera azken analisiaren parte ez direnak ere sortu egiten direlako.

Adibidez, lexikoan *mendi* lemarako kategoria bakarrik badugu definituta (alegia, izena dela), eta izen guztiek "izen + adjektibo -> IM" erregela baten parte izateko aukera dutela esaten badugu, *mendi* eta edozein adjektiboz osatutako segida bat onartuko dugu. Alegia: *mendi* (izena; hiztegian, bere deklinabide-aukera guztiekin); izen + adjektibo -> IM (erregela sintaktikoa). Era berean, IM + det -> IS moduko beste erregela bat badugu, *mendi* izenari aplikatuz, edozein adjektibo eta determinatzailerekin agertuko zaigu. Erregela horiekin, ordea, *mendi bat, mendiak* edo *mendi zenbait* onartzen ditugun bezalaxe, *mendiak zenbait* modukoak ere onartzen ditugu.

Beraz, hori ekiditeko alderdi murriztailea beharrezkoa da. Esan dugun bezala, sistema batzuek alderdi hori ez dute gehiegi lantzen, eta eraikitzea da axola zaiena; beste sistema batzuek, berriz, murrizketa dute zeregin nagusia.

Murrizketa hori lexikoko informazioa zehaztuz edo erregela sintaktikoetan bertan deskribatuz egin daiteke. Esaterako, *etxeak zenbait* katea okertzat ematea nahi badugu, hainbat aukera egongo dira:

- *Zenbait* determinatzaileari lexikoan esplizituki markatzea mugagabea dela, eta ondoren sintaxiak saihets ditzala mugatu pluraltasuna eta mugagabetasuna IS berean (ekuazioen bidez definitutako printzipioetan oinarritua)

- Baldintza horiek erregela sintaktikoetan definitzea. Alegia, IM + det -> IS ontzat eman, baina gero, beste erregela baten bidez, adierazi determinatzaile gisa *zenbait* dugunean, ondoko izenaren deklinabide-atzizkia mugagabea izan behar duela.

Hitz bakarreko adibide batera joaz, eta adibidez IXA taldearen analizatzaile morfologikoak ematen dituen (eraikitzen dituen) analisi guztiak kontuan izanda, hauxe da *mendiak* hitzerako dugun analisi morfosintaktikoa:

mendiak
lema: mendi IZE ARR
morf: ak DEK, ABS, PL, MUG, FS1 @OBJ / FS2 @SUBJ / FS3 @PRED
EDO
lema "mendi IZE ARR
morf: ak DEK ERG S M FS1 @SUBJ

Analisi posibleek erakusten digutenez, *mendiak* forma aditzaren subjektu (ergatibo singularra), objektu (absolutibo plurala) edo elementu predikatiboa (absolutibo plurala) izan daiteke:

Mendiak hirietatik oso hurbil daude -> subjektua

Opor hauetan Italiako mendiak ezagutu ditugu -> objektua

Alpeak mendiak dira -> predikatiboa

Sistema murriztailearen zeregina, hortaz, aukera hori egitea (desanbiguatzea, azken batean) da.

3.2.4 Analisi sintaktikorako (*parsing*) teknikak: *top-down* eta *botton-up*

Behin gramatika erregelez osatua dela, analisi sintaktikoa goitik behera (*top-down*) edo behetik gora (*botton-up*) egin daiteke. Jo dezagun, esaterako, ondoko erregelak ditugula:

P: IS + AS

IS: IM + Det

IM: I + adj

AS: A

Eta eman dezagun izen, adjektibo, aditz eta determinatzaileen hiztegia dugula: Hau da:

Adjektiboa: edozein adjektibo (*gorri, zuri, handi, polit*…)

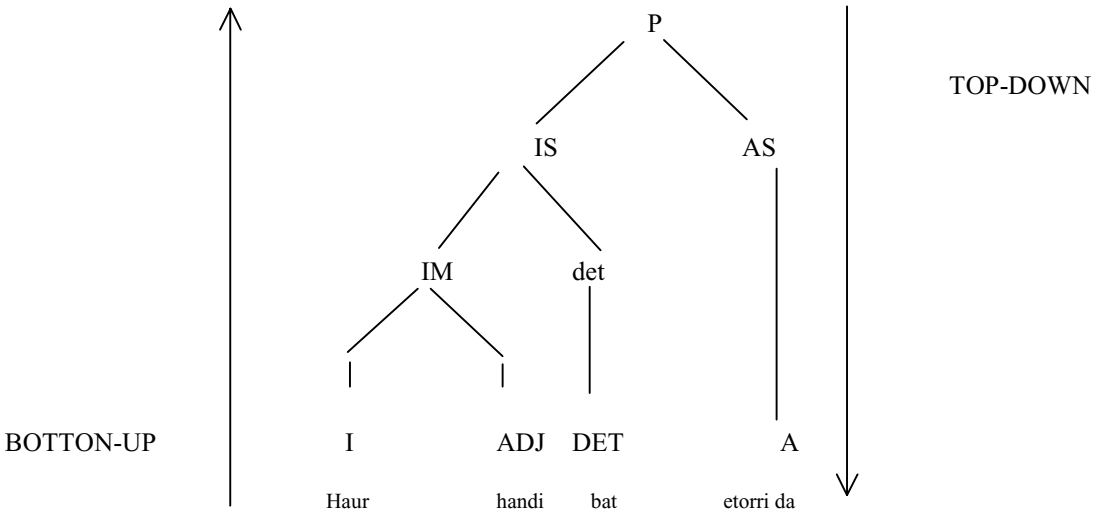
Izena: edozein izen (*etxe, gizon, egur, begi, ume*…)

Aditza: edozein aditz (*ekarri, joan, izan, egin*…)

Determinatzailea: edozein determinatzaile (*zenbait, batzuk, bat*…)

Hori izanda, guk zer aztertu nahi dugun, goitik beherako edo behetik gorako teknika erabiliko dugu.

Goitik beherako teknika erabiliz, unitate handitik abiatzen gara, alegia, P-tik, eta gero zuhaitzean behera "hostoetaraino" iristen gara, hau da, elementu terminaleraino. Behetik gorakoan, hostoetatik hasi, eta, eskuin-ezkerrera elementuak aurkitu ahala, unitate orokorragoak osatuz doaz.

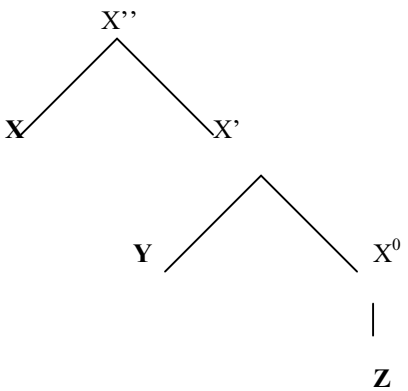


Sintaxi partziala egiten denean, arruntena *bottom-up* teknika erabiltzea da, unitate orokorrak (P, esaterako) gutxiago aztertzen direlako ikerketa desberdinetan (adibidez, aditz bat eta inguruan agertzen diren sintagmak ateratzeko).

3.2.5 Erlazio sintaktikoen adierazpidea: zuhaitz-egiturak, osagai-egitura, mendekotasun-egitura

Esan dugun bezala, erlazio horiek adierazi egin behar dira, baina horretarako ere hainbat adierazpide daude.

Ikuspegi teorikoagoa duten formalismoek, esan dugun bezala, ez dituzte erlazio guzti-guztiak (luzeak, anbiguoak, okerrak) aurreikusten, eta proposatzen dituzten analisiak esaldi motzagoetara mugatzen dira. Ikuspegi teoriko gehienek zuhaitz-egiturak erabiltzen dituzte. Esaterako, ezagunena, Chomsky aitzindari duen korronte sortzailetik datorren X-barra teoriako adierazpidea dugu:



Teoria horren arabera, kategoria guztiek egitura honetan parte hartzen dute beren sintagma-izaera jasotzeko. Hori dela eta, kategoria jakin bateko hitzetik abiatuta, horrelako egitura sintaktikora iristen da.

Ordea, aplikazio praktikoetara zuzenduta dauden formalismoak, azaleko sintaxian dagoen guztia ezagutu behar dutenez, hasteko egitura lauetatik abiatzen dira, eta gero hortik abiatuta zuhaitzak eraikitzen saiatzen dira. Etiketatze sintaktiko hauen bidez, hain zuzen, *Treebank* deituriko corpus sintaktikoki etiketatuak lortzen dira. Adierazpide ezagunenen artean *osagai-egitura* eta *mendekotasun-egitura* ditugu. Lehenak parentesiak erabiltzen ditu osagaiak lotzeko eta bereizteko, eta bigarrenak erlazio horiek esplizitu adierazten ditu dependentzia-etiketen bidez. Adib.:

Osagai-egitura: (P (IS X) (AS (IS Y) (A Z)))

Mendekotasun-egitura: ncsubj[14] (- , Z, X)
ncobj[15] (- , Z, Y)

Argi dago, hortaz, formalismoak, eta hauek erabiltzen dituzten gramatikak, desberdin samarrak izango direla hautatutako ikuspegi eta irizpideen arabera.

Ikus ditzagun segidan sarreran aipatutako hurbilpen motak, eta horietan oinarritutako sistema batzuk.

3.3 Hurbilpenak

Oro har, hiru hurbilpen mota daude sintaxi konputazionala lantzeko. Batetik, ezagutza linguistikoan oinarritutakoa; bestetik, teknika probabilistikoetan oinarritutakoa; eta azkenik, biak, ezagutza linguistikoa eta teknika probabilistikoak konbinatzen dituena. Gehien erabiltzen direnak ezagutza linguistikoan oinarritutako sistemak dira, eta horien barruan bi joera daude: testuingururik gabeko gramatiketan oinarritutako sistemak (TGG sistemak, alegia); eta egoera finituko mekanismoetan oinarritutako sistemak. Ondorengo lerroetan horiek guztiak izango ditugu aipagai, baina batez ere, ezagutza linguistikoan oinarritutako sistemetan sakonduko dugu.

3.3.1 Ezagutza linguistikoan oinarritutako sintaxia

3.3.1.1 Testuingururik gabeko gramatiketan oinarritutako sistemak (TGG sistemak)

Testuingururik gabeko gramatiketan oinarritutako sistemak hizkuntzaren teoriaren ikuspegitik eraikiak izan ohi dira. Horrenbestez, hauetako askoren helburu nagusia hizkuntzaren teoria garatzea izan da, eta hizkuntza analizatzeko tresnen eraikuntza bigarren mailan utzi da. Dena dela, kasu batzuetan teoria eta aplikagarritasuna lotzeko ahalegin berezia egin da, sistema eraginkorra lortzeko asmoarekin.

---

[14] ncsubj: "*non-clausal subject*"

[15] ncobj: "*non-clausal object*"

Gramatika hauek esaldien egitura hierarkikoa eta errekurtsibitatea definitzeko dira egokiak. Hori lortzeko erabiltzen diren mekanismoak, halere, sinpleagoak edo konplexuagoak dira. Batetik TGG sinpleak izango ditugu eta bestetik *Baterakuntza*n oinarritutakoak.

TGG sinpleetan erregelek osagai atomikoak besterik ez dute deskribatzen. Horren ondorioz, egitura sintaktiko sinpleenak (adibidez, izen-sintagma baten komunztadura kontuan hartzea) zehazki definitzeko dozenaka erregela behar dira. Hori ez gertatzeko, gramatikaren osagaiei informazioa gehi dakieke ezaugarri-egituren bitartez (Shieber, 1986), horrela gramatikaren trinkotasuna eta sinpletasuna bultzatuz. Informazio linguistiko hori erabiltzeko baterakuntza izaten da eragiketarik inportanteena. Irabazpen honen kontra baterakuntza-ekuazioen kalkuluaren denbora-kostua dugu, eraginkortasuna moteldu egiten baita. Ikus euskarako adibide bat: izen + adjektibo segida batek IS1 delako sintagma osatzeko erabiltzen den erregela; erregela honetan izen bereziak (pertsona nahiz lekuzkoak) ez onartzea ere zehazten da, bai eta adjektiboaz (izenondoaz) gain aditzetiko partizipioa onartzea:

Erregela: **is1 --> ize + adj**
adibidea: *gizon handi*

```
rule (r_is1_2, X0 ---> [X1, X2]@[
                    m(1,  eta ([X1/kat   <=>    ize,
                                X1/gunelex/nag/azp ez [izb, lib]])),
                    m(2,  edo ([eta([X2/kat              <=>    adj,
                                    X2/gunelex/nag/azp   badago   [izo]]),
                                eta [X2/kat              <=>    adi,
                                    X2/sint/nag/adm      badago [part],
                                ])),
```

Sistema hauetan, hortaz, lexikoaren eta gramatikaren artean oreka desberdina egongo da. Zenbaitzuek lexikoan kodetuko dute sintaxian kodetu ohi den informazioa, eta beste batzuek erregela sintaktikoetan definituko dituzte operazio lexikalak. Baterakuntza oinarritzat duten formalismorik ezagunenak, *Lexical Functional Grammar* (LFG; Bresnan, 1982), *Generalized Phrase Structure Grammar* (GPSG) eta *Head-Driven Phrase Structure Grammar* (HPSG) ditugu.

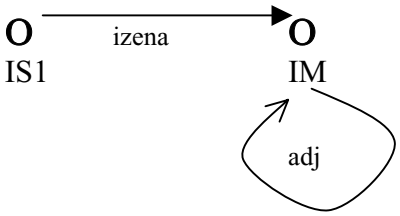
3.3.1.2 Egoera finituko mekanismoetan oinarritutako sintaxia

Egoera finituko mekanismoen artean egora finituko automatak eta transduktoreak ditugu (Roche eta Schabes, 1997). Mekanismo hauek oso erabiliak izan dira informatikako alor askotan, baina orain dela gutxira arte ez dira egokitzat hartu linguistika konputazionalaren eginkizun nagusietan: hiztegien kodeketan, testuen prozesamenduan eta hizketaren prozesamenduan. Hala ere, azken urteotan egindako lanen ondorioz, esan daiteke teknika hauek etorkizun oparoa izan dezaketela lehenago TGGen bidez egiten ziren eginkizun askotan, beraien inplementazio eraginkorrei esker.

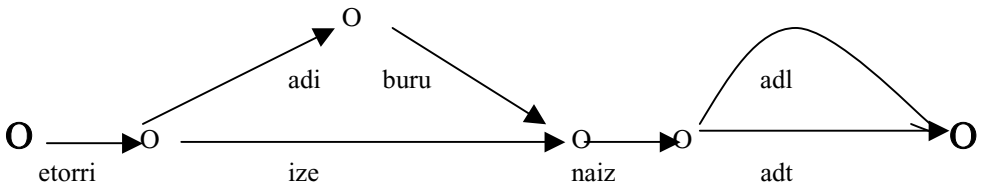
Egoera finituko automatak eta transduktoreak dira hurbilpen honen funtsa, batzuk lengoaiak eta besteak lengoaien arteko erlazioak definitzeko. Horien zehaztapena egiteko adierazpen erregularren lengoaiak erabiltzen dira. Sistema horien ezaugarri nagusienak homogeneotasuna, malgutasuna eta modulartasuna

dira. Horretaz gain, automaten propietate algoritmikoei esker (determinista bihurtzeko eta minimizatzeko garrantzitsuenak), gramatikak era trinkoan gorde daitezke, eta modu eraginkorrean aplikatu ere.

Ikus dezagun esandakoa garbiago adibideen gainean. Har dezagun berriro *gizon handi* kate-segida, IS1 deituriko sintagma maila osatzen duena. Hori honelaxe osatuko litzateke egoera finituko mekanismoen bidez:

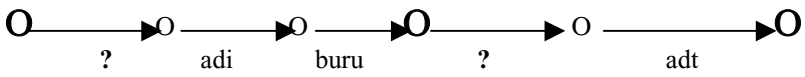


Plantea dezagun adibide konplexuago bat. Eman dezagun bi hitzek osatutako esaldi sinple bat aztertu behar dugula eta horietako hitz bakoitzak bi interpretazio posible dituela (EDBLn hitzok duten informazio morfologikoan oinarritzen gara hemen; alegia, horixe da erabiliko dugun iturri lexikala): “etorri naiz”: *etorri* (izena / aditza forma burutuan), *naiz* (aditz laguntzailea / aditz trinkoa). Guztira lau bide daude automata hori zeharkatzeko, hau da, lau interpretazio dauzka anbiguoa den esaldi horrek. Horrela errepresentatuko genuke automata modu sinplifikatuan:

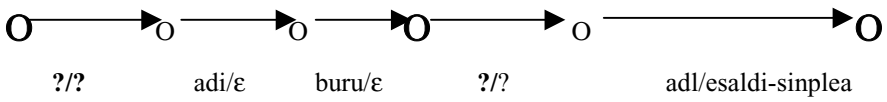


Lau interpretazioak: adi buru + adl; adi buru + adt; ize + adl; ize + adt.

Automata horren aurrean, orduan, alderdi murriztailea jarri beharko da martxan, eta hori bi modutara egin daiteke: ezinezko testuinguruak debekatuz edota testuinguru zuzenak aukeratuz. Bi murriztapen mota horiek osagarriak direnez, askotan gramatikariaren gustuaren arabera erabiltzen da bata edo bestea. Goiko adibideari ondorengo debeku bidezko automata aplikatuz gero, hiru interpretazio geratuko zaizkigu: adi buru + adl; ize + adl; ize + adt.



Ikuspegi eraikitzailean oinarritutako formalismoetan, berriz, esaldi baten hasierako informazioa hartuta emaitza gisa lortu nahi diren egiturak eraiki edo sortzen dira. Hori egiteko transduktore (*trasducer*) izeneko automata bereziak erabiltzen dira. Esaterako, eta adibide berak gogoan:



Arku bakoitzean jarri ditugun bi ikurrek sarrerako eta irteerako lengoaiak definitzen dituzte. Automata horren lana, hortaz, sarrerako lengoaiaren instantziak irteerako lengoaiarekin ordeztea izango da. ε ikurrak kate hutsa adierazten du, hau da, beraren bitartez nahi ez diren ikurrak irteeratik ezabatzen dira. Adibide gisa jarritako transduktore honek aditz burutua eta aditz laguntzailea aurkitzean esaldi sinple baten arkua gehituko du, 'etorri + naiz' segida esaldi sinple gisa analizatzeko.

3.3.2 Teknika probabilistikoetan oinarritutako sintaxia

Hurbilpen probabilistikoa (Black *et al.,* 1993) indar handikoa bihurtu da azken hamarkadan, aurreko lanetan gramatikariek egiten zituzten atazak automatikoki egiteko. Sistema hauen ezaugarri nagusien artean hauek ditugu:

- Corpus etiketatuen (*Treebank*) beharra. Analizatzaile mota hauetan lan gehiena corpus etiketatuetatik ateratako probabilitateen bidez egiten da, hau da, gramatikak garatzeko orduan eskuzko lan gramatikal minimoa egiten da, ezagumendu linguistikoa corpusean agertzen diren elementuetatik (eta beren maiztasunetatik) ateratzen baita. Corpus horiek ingeleserako landu dira gehienbat (*Brown Corpus, Penn Treebank* (Marcus eta Santorini, 1991)). Adibidez *tagger* edo etiketatzaileetan probabilitate lexikalak kategoria edo hitzen bigrama edo trigramen bidez ateratzen dira (Church, 1998; Garside *et al.,* 1987; Brill, 1995; Charniak, 1993), gero testu berrietan aplikatzeko. Esan behar da corpus bat etiketatzea lan luzea eta zaila dela. Gainera, gertaera linguistikoen deskribapen zabala izateko, corpusak tamaina handia izan behar du, fenomeno askoren agerpen-maiztasuna txikiegia gerta ez dadin.
- Azaleko sintaxia. Sistema probabilistiko gehienetan azaleko analisia egiten da; etiketatzaileetan adibidez, hitz bakoitzaren kategoria sintaktikoa igarri behar da. Nahiz eta egitura sintaktiko osoak lortzeko zenbait lan egin (Atwell, 1987; Bod 1993), orandik frogatzeke dago estatistika hutsean oinarritzen den analizatzaileen bideragarritasuna.
- Muga gaindiezinak. Sistema hauek orain arte gaindiezinak izan diren mugak dauzkate. Adibidez, etiketatzaile estatistikoetan, % 95-97 inguruko neurriak (Voutilainen, 1994a; Brill eta Wu, 1998) agertu dira zenbait lengoaiatarako. Nahiz eta neurri hori ona izan lehenagoko etiketatzaileekin konparatuz gero, horrek problema bat supostuko du edozein analizatzailerentzat, zenbaki hori muga maximotzat onartuko bagenu, esaldi askotan errore bat egotea suposatuko lukeelako.

3.3.3 Teknika linguistiko eta probabilistikoen konbinazioa

Teknika probabilistikoetan oinarritutako sistemetan ikusi ditugun mugak kontuan izanda, ezin izango dira espero emaitza hain onak sintaxi osoaren tratamendu probabilistikoan. Bestalde, linguistek idatzitako gramatiketan, maila altuko gertaera linguistikoak deskribatu dira gehienbat, sintagmak zein esaldi osoak konbinatzeko, baina arreta gutxiago eskaini zaio esaldi errealetan agertzen den zenbait fenomenori, egitura baten maiztasuna kasu. Horregatik, metodo probabilistikoak eta ezagutza linguistikoa lotzeko saioak egin dira, bakoitzaren abantailak biltzeko asmoz.

Adibidez, (Black *et al.,* 1993) lanean hizkuntzalariek egindako gramatika bat erabiltzen da, baina erregelen aplikazioa sintaktikoki etiketatutako corpus batetik ateratako probabilitateen bidez erabakitzen da. Hasiera batean, analizatzaileak aukera posible guztiak proposatzen ditu, nahiz eta batzuk probabilitate gutxikoak izan, ondoren probabilitate handienekoa aukeratzeko. Horrela, analizatzailearen lana erregela horietatik abiatuta, corpusaren probabilitateetatik hurbilago dagoen analisia ateratzen da, eta emaitza zuzena emango da baldin eta analizatutako esaldiaren egitura bat badator corpusean dauden erregelen aplikazioen maiztasunarekin. Horrelako saioek erakusten dute sintaxiaren eredu eraikitzaileen (baterakuntzan oinarritutako testuingururik gabeko gramatika) eta murriztaileen (corpusen probabilitateen bidez probabilitate handieneko analisia aukeratzeko) beharra.

Bod & Kaplan-ek (1998) azaltzen duten sisteman antzeko bilketa egitea proposatu da. Abiapuntua LFG gramatika bat eta corpus batetik ateratako egitura sintaktikoak ditugu. Aurrekoarekiko desberdintasun nagusia da, honetan, egituren maiztasunak neurtzen direla, eta besteetan, erregela sintaktikoen aplikazioen maiztasunak.

Aipatutako sistemek derrigorrean eskatzen dute corpus etiketatua, eta hori muga bat da euskara bezalako hizkuntzentzat. Muga hori gainditzeko, Carrol & Rooth-en lanean (1998) etiketatu gabeko sistema bat deskribatzen da. Bertan, oinarrizko gramatika bat erabiliz, corpus baten azterketarako gramatika probabilistikoa lortzen dute, EM (*Expectation-Maximization*) algoritmoa eta gramatika probabilistiko lexikalizatuak (PLCFG, *Probabilistic Lexikalized Context-Free Grammar*) erabiliz. Gainera, hasierako gramatika lexikalizatu egiten dute, hau da, hasierako erregela sintaktikoei dagozkien hitzen probabilitateak gehitzen dizkiete, horrela corpus desberdinetarako gramatika egokitzeko aukera emanez, ikaste-prozesu baten ondoren. Bide hau interes handikoa da euskararen kasuan, etiketatze-lana ekiditen duelako. Dena dela, oraindik metodo honen lehen esperimentuak egiten ari dira, emaitza onekin, eta ikusi beharko da gramatika zabal eta lexikoi handien erabilerarekin mantentzen diren, une honetan dagoen konputazio-baliabideen arazoa (denbora eta espazioa, eredu probabilistiko konplexuen ondorioz) gainditzuz.

3.4 Egungo egoera

Testuingururik gabeko formalismoetan oinarritutako sistema inplementatuen artean, hauexek ditugu, besteak beste: *Alvey Natural Language Tools* (ANLT) (Carrol, 1993; Grover *et al.,* 1993); *Programming Language for Natural Language Processing* (PLNLP sistema) (Jensen *et al.,* 1993); TACAT (Atserias *et al.,* 1998); *Freeling* espainerararko analizatzailea (Carreras et al., 2004); *Core Language Engine* (Alshawi & Moore, 1992); *Government and Binding* (GB) (Berwick *et al.,* 1991); PC-PATR (Antworth, 1994); *Categorial Grammar* (Uszkoreit, 1986).

Murriztapen Gramatika (MG, Murriztapen Gramatika) (Voutilainen eta Tapanainen, 1993; Karlsson *et al.,* 1994, Voutilainen, 1994ab) formalismoa ikuspuntu murriztailetik definitu da, eta azken urteotan azaleko sintaxia eta desanbiguazioa lortzeko egindako sistemetatik arrakastatsuenetako bat bihurtu da, oso ezaugarri desberdinetako hizkuntzetara aplikatu delako. Esaldi bat emanda, lehen pausoa hitz-forma bakoitzari etiketa morfosintaktiko posible guztiak gehitzea da. Ondoren, murriztapen-erregelen multzoa aplikatzearen ondorioz, hitz-forma bakoitzak interpretazio bakarra eta zuzena izatea bilatzen da. Murriztapen-erregela horiek arestian aurkeztu ditugun debekuen modukoak edo interpretazio zuzenak aukeratzekoak dira. Emaitzan, beraz, hitz bakoitza morfosintaktikoki desanbiguatuta egongo da.

MG formalismoak, batez ere murriztailea izan arren, badu aukera bestelako etiketak, alegia, etiketa berriak, esleitzekoa. Ahalmen horri esker, informazio sintaktiko gehiago eransten da hitzetan (esaterako, sintagma-hasierako eta -bukaerako markak, sintagmen gune-marka), eta horien bidez zuhaitz sintaktikoak eraikitzeko bidea irekitzen da (Koskenniemi *et al.,* 1992; Voutilainen, 1994b).

Azkenik, ildo beretik, Xerox-eko ikerketa-taldeak adierazpen erregularren bidezko tresna linguistikoak , XFST (*Xerox Finite State Tool*) izenekoa, garatu dituela aipatuko dugu (Karttunen *et al.,* 1997; Aït-Mokhtar eta Chanod, 1997; Chanod eta Tapanainen, 1996ab). Adierazpen erregularrak analisi morfologikorako erabiltzen ziren orain dela gutxira arte, baina dagoeneko beste lan batzuetan aplikatu izan dira, tokenizaziotik hasita azaleko analisi sintaktikora arte. Tresna hauetan ikuspegi eraikitzailea eta murriztailea erabiltzea dago. Lortutako tresnek malgutasuna, adierazpen erregularren sinpletasuna eta teoria matematiko baten sendotasuna dituzte ezaugarri nagusi aipagarritzat. Guk, hala ere, ez dugu sakonduko tresna hauetan.

Euskaraz, sintaxi partzialari ekiteko, bi bide hautatu dira: bata Murriztapen Gramatika (MG) —ezagutza linguistikoan oinarritutako egoera finituko analizatzailea—, eta PATR II —ezagutza linguistikoan eta baterakuntzan oinarritutako analizatzailea—. Hona hemen bakoitzaren zertzelada batzuk.

Murriztapen Gramatika (MG)

*Murriztapen Gramatika*ren helburua ez da, beste kasu askotan gertatu ohi den legez, "jostailuzko" gramatika bat egitea laboratorioko esaldiekin jolastu ahal izateko. MG **benetako testuak analizatzeko** pentsatuta dago; testu gordinekin erabiltzeko, alegia. Horrela izanik, **beste aplikazio batzuetarako oinarri** gisa balioko du.

Bestalde, MGk morfologian oinarritutako ***parsing***a bultzatzen du, eta esan beharrik ez dago hori zein ongi egokitzen zaien euskara bezalako hizkuntzei, non morfologia eta sintaxia hain loturik dauden. MGren bidezko analisiaren parte inportanteenetako bat **desanbiguazio morfologikoa** da, hau da, analisi morfologikotik irtendako emaitza anbiguoa tratatzea, ezagutza linguistikoan oinarritutako **murriztapen-erregelen bidez.**

Formalismoa egokia da euskararen tratamendu sintaktiko orokorrean erabiltzeko, zeren beste formalismo batzuek ezartzen duten zurruntasunetik urrun, eta arazoak arazo, perpaus barneko elementuen ordena librea-edo duten hizkuntzentzat egokia delakoan baikaude.

Orain arte eta Euskararako Lematizatzaileari (EUSLEM) begira, **anbiguotasun kategoriala** ebazteari eman zaio lehentasuna.

1. adibidean *Gero hegoak moztu eta pospolo-kaxa batean gartzelaratuko zizkizun.* esaldiaren analisia daukagu. Ikus daitekeenez, hitz bakoitzeko analisi posible bat baino gehiago ematen da.

MGko erregela batzuk aplikatu eta gero, hitz bakoitzeko analisi bakarra uzten saiatzen da. Ikus 2. adibidean nola geratzen den analisia disanbiguazio-erregelak aplikatu ondoren.

Esate baterako, 187. erregelan **(@w =! ETOR (0 C PART) (NOT 1 DET))** hauxe adierazten da:

- **@w =! ETOR**
  ETORkizuneko ezaugarria hartu (=!)
- **(0 C PART)**
  baldin eta hitz horren interpretazio guztiak PARTizipioak badira
- **(NOT 1 DET)**
  eta urrats batera eskuinetara ez dagoen DET ezaugarririk.

Gaur egun MG beste zereginetarako ere erabiltzen da, hala nola, sintagma-hasierak eta sintagma-bukaerak, eta perpaus-mugak markatzeko.

```
  "<$.>"
  PUNT_PUNT
"<Gero>"
  "gero"  ADB ADO  HAS_MAI @ADLG
  "gero"  IZE ARR  DEK ABS MG @OBJ @SUBJ  HAS_MAI
  "gero"  IZE ARR  ZERO HAS_MAI @KM>
"<,>"
  PUNT_KOMA
"<hegoak>"
  "hego"  IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ
  "hego"  IZE ARR DEK ERG NUMS MUGM @SUBJ
"<moztu>"
  "motz"  ADI SIN ASP PART DEK ABS MG @OBJ @SUBJ
  "motz"  ADI SIN ASP PART  ZERO NOTDEK @-JADNAG
"<eta>"
  "eta"  LOT JNT @PJ @SJ AORG
  "eta"  LOT MEN KAUS @MP AORG
"<pospolo>"
  "pospolo"  IZE ARR DEK ABS MG @OBJ @SUBJ
  "pospolo"  IZE ARR  ZERO @KM>
"<kaxa>"
  "kaxa"  IZE ARR DEK ABS MG @OBJ @SUBJ  AORG
  "kaxa"  IZE ARR DEK ABS NUMS MUGM @OBJ @SUBJ  AORG
  "kaxa"  IZE ARR  ZERO AORG @KM>
"<batean>"
  "bat"  DET DZH DEK NUMS MUGM DEK INE @ADLG
  "bat"  IZE ARR DEK NUMS MUGM DEK INE @ADLG
  "bate"  IZE ARR DEK NUMS MUGM DEK INE @ADLG
"<gartzelaratuko>"
  "gartzelara"  ADI SIN ASP PART ASP ETOR  NOTDEK AORG
  "gartzelara"  ADI SIN ASP PART DEK NUMS MUGM DEK GEL @IZLG> @<IZLG
                @ADLG DEK ABS MG @OBJ @SUBJ  AORG
  "gartzelara"  ADI SIN ASP PART DEK NUMS MUGM DEK GEL @IZLG> @<IZLG
                @ADLG  AORG @-JADNAG
"<zizkizun>"
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU LOT MEN @+JADNAG_MP @+JADLAG_MP
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU LOT MEN ERLT @+JADNAG_IZLG>
           @+JADLAG_IZLG>
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU @+JADLAG
"<$.>"
  PUNT_PUNT
```

1. adibidea: *Gero hegoak moztu eta pospolo kaxa batean gartzelaratuko zizkizun.* esaldiaren analisia

```
"<$.>"
  PUNT_PUNT
"<Gero>" D:395
  "gero"  ADB ADO  HAS_MAI @ADLG
"<,>"
  PUNT_KOMA
"<hegoak>" D:223
  "hego"  IZE ARR DEK ABS NUMP MUGM @OBJ @SUBJ
"<moztu>" D:16
  "motz"  ADI SIN ASP PART  ZERO NOTDEK @-JADNAG
"<eta>" D:392
  "eta"  LOT JNT @PJ @SJ AORG
"<pospolo>"
  "pospolo"  IZE ARR DEK ABS MG @OBJ @SUBJ
  "pospolo"  IZE ARR  ZERO @KM
"<kaxa>" D:30
  "kaxa"  IZE ARR  ZERO AORG @KM
"<batean>" D:164
  "bat"  DET DZH DEK NUMS MUGM DEK INE @ADLG
"<gartzelaratuko>" D:187
  "gartzelara"  ADI SIN ASP PART ASP ETOR  NOTDEK AORG @-JADNAG
"<zizkizun>" D:208
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU LOT MEN @+JADNAG_MP @+JADLAG_MP
  "*edun"  ADL B1 NR_HK NI_ZU NK_HU @+JADLAG
"<$.>"
  PUNT_PUNT
```

2. adibidea: 1. adibideko esaldiaren analisi desanbiguatua

PATR-IXA

Esana dugun bezala, PATR baterakuntzan oinarritutako analizatzaile sintaktiko malgua da. Hala, ezaugarri-egiturekin lan egiten du eta ezaugarri horiek erregela sintaktikoen bidez konbinatzen dira.

Gramatikaren azalpenarekin hasteko, ikus dezagun lehenago nolako izen-sintagma edota adizlagunak onartzen dituen gramatikak, eta geroago aztertuko dugu perpaus mailakoa. Izen-sintagmaren mailan hiru eraketa posible bereizi dira:

1 Buru gisa izen arrunta dutenak. Azkeneko hitzean deklinabide-atzizkia etortzen da beti (*knmdek* izena eman zaio zati horri). Izenaren aurretik izenlagun bat (*izlg*) edota determinatzaile bat (*det*) aukerakoak dira. Atzetik adjektibo bat (*adj*) edota determinatzaile bat (*det*) aukerakoak dira. Ekuazioen bidez kontrolatzen da zer determinatzaile etor daitekeen aurretik eta zein atzetik.

(izlg) +	**(det) +**	**ize +**	**(adj) +**	**(det) +**	**knmdek**
etxeko	*altzari*	*zahar*		*hori___ekin*	
etxeko lau	*altzari*	*zahar_____etan*			
etxeko		*altzari*	*zahar_____ari buruz*		

2 Buru gisa izen berezia dutenak. Izen bereziaren aurretik izenlagun bat aukerakoa da, baina ez da onartzen determinatzailerik, ezta adjektiborik ere eraketa mota honetan.

(izlg) +	**izb**	**+**	**knmdek**
Donostiako	*Peru_____ri*		

3 Buru gisa izenordea dutenak. Deklinabide-atzizkia baino ez da onartzen kasu honetan.

ior	+	**knmdek**
ni_____		_____ri

Horrelako egiturak onartu ahal izateko, osaketa sinpleenetik hasi eta konplexuenera heltzeko, *is1*, *is2* eta *is3* kategoria laguntzaileak bereizten dira. Osaketa konplexuena biltzen duen *is3* motako osagai bati deklinabide-atzizkia (*knmdek*) lotuz *isk* lortzen da. Definitzen diren *isk* horiek ez dira izen-sintagmak bakarrik, izen multzo bati edozein deklinabide-atzizki erantsita lortzen diren guztiak baizik. Beraz, *isk* horiek batzuetan izen-sintagma eta beste batzuetan adizlagun ditugu, beraien arteko diferentzia nagusia kasua delarik.

Deklinabide-atzizkia modu orokorrean hartzen da. Kasu, numero eta mugatasunaren informazioak dakartzan atzizki multzoari, esana dugun bezala, *knmdek* izena eman zaio. Are gehiago, postposizio kasuetan (adibidez *-ri buruz* edo antzekoak erabiltzen ditugunean) atzizkia bera gehi beste hitz bat ere hartzen du *knmdek* delako horrek.

Analizatzen diren izen-sintagmen osaketa ondoko erregela[16] hauetan ikus daiteke:

is1 →	*ize adj*	*etxe EDER*
	/ *ize*	*etxe*
is2 →	*det is1*	*ZENBAIT etxe eder*
	/ *is1 det*	*etxe eder BAT*
	/ *is1*	*etxe*
	/ *izb*	*JON*
is3 →	*izlg is2*	*MENDI HORRETAKO zenbait etxe eder*
	/ *is2*	*zenbait etxe eder*
	/ *ior*	*ZU*
isk →	*is3 knmdek*	*etxe ederrEKIN*
		mendiko zenbait etxe ederrEK
		mendiko zenbait etxeRI BURUZ

Izen-sintagmaren deskribapena bukatzeko esan dezagun izenlagunaren egitura nagusiak *isk* osagaien egitura bera duela, baina kasua genitiboetako bat izan beharko dela:

izlg → is3 + knmdek(gen/gel)	*mendi horretaKO*
izlg →as + erlt	*nik ikusi duDAN*

Perpausa analizatzeko orduan subjektua ez da bereizten, eta horrela perpausa aditz-sintagma gisa hartzen da beti. Aditz-sintagma sinpleena aditza besterik ez duena da; aditz trinkoa zein aditz-erroa gehi laguntzailea, aukera biak onartzen dira. Aditza ezagutu ondoren, aditzaren ezker zein eskuinaldean ager daitezkeen osagaiak banan-banan onartzen dira, segidan azalduko ditugun erregelak erabiliz:

1 Kasu nuklearrak analizatzeko erregelak. Ergatibo, absolutibo edo datibo kasuak hartzeko erregela hauetan, ekuazioen bidez, numero, kasu eta pertsona ezaugarrien komunztadura

[16] Benetako erregelen sinplifikazioa erakusten dugu hemen. Hauek ez dira gramatikako erregelak osotasunean, horiek murriztapenak adierazteko eta osagai sintaktiko berriak sortzeko ekuazio multzoa baitute.

egiaztatzen da. Horrela, esate baterako, '**Peruk txakurrak ekarri du*' bezalako esaldiak ez dira onartuko. Erregelak bikoiztuta daude kasu horiek aditzaren aurretik zein atzetik onartu ahal izateko:

as→	*isk(erg) as*	*GIZONEK ikusi dute*
as→	*isk(abs) as*	*GIZONAK ikusi dituzte*
as→	*isk(dat) as*	*GIZONARI eman dio*
as →	*as isk(erg)*	*ikusi dute GIZONEK*
as→	*as isk(abs)*	*ikusi dituzte GIZONAK*
as→	*as isk(dat)*	*eman dio GIZONARI*

2 Adjuntuak tratatzeko erregelak. Ergatibo, absolutibo edo datibo ez diren kasuak hartzeko erregelak dira hauek. Hemen eta hurrengo puntuetan ez ditugu erregelak era bikoiztuan erakutsiko, baina suposatu behar da elementua aditzaren atzetik ere onartzeko beste erregela bat definitu dela:

as→	*isk(ez da abs, erg edo, dat)*	*as*	*GIZON HORREKIN ikusi dute*

3 Adberbioak tratatzeko erregelak:

as→	*adb as*	*GAUR egin dut*

4 Mendeko perpausak tratatzeko erregelak: konpletiboak, zehargalderak, moduzkoak eta denborazkoak:

as →	*mend-modu-denb as*	*HONA NENTORRELA ikusi dut*
as →	*mend-zehargaldera as*	*EA JOAN DEN galdetu du*
as →	*mend-konp as*	*ETORRI DIRELA jakin da*

5 Guztira 90 erregela definitu dira eta batez beste 15 ekuazio ditu erregela bakoitzak. Esan bezala, erregelak hemen aurkeztu baino konplexuagoak dira, bai notazio aldetik, bai ekuazioen aldetik, bai ñabarduren aldetik. Erregela baten (`is3 → izlg + is2`) benetako itxura ikus daiteke hemen:

```
X0 ---> X1, X2
        X1 kat              = izlg
        X2 kat              = is2
        X0 kat              = is3
        X0 sint kom         = X2 sint kom
        X0 gunelex          = X2 gunelex
        X0 sint osgk izlg   = X1 sint izlg
        X0 sint osgk adj    = X2 sint osgk adj
        X0 sint osgk detaur = X2 sint osgk detaur
        X0 sint osgk detatz = X2 sint osgk detatz
        X0 sint nag kom     = X2 sint nag kom
        edo[eta[ X2 gunelex nag kat = eli,
                 X0 forma = X1 forma],
                 X0 forma = $($(X1 forma, "+"),
                              X2 forma)]
```

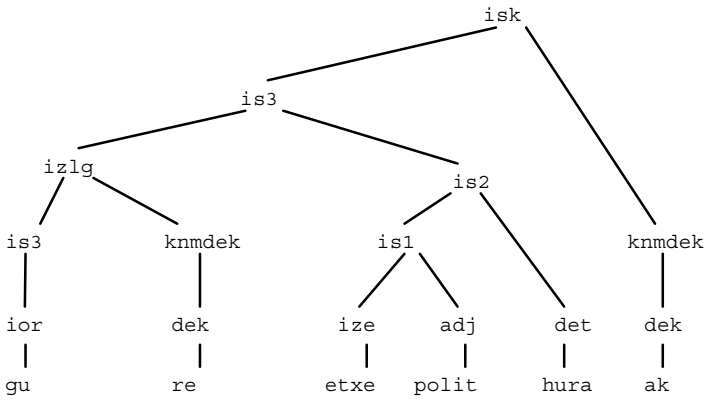
Zer analiza dezake gramatika honek? Zein da bere estaldura? Maila lexikoan oso estaldura handia dagoenez (esan dugun bezala, EDBLn 83.070 sarrera erabiltzen dira analisian), esan dezakegu testu errealetako ia izen-sintagma eta adizlagun guztiak analiza daitezkeela. Berdin esan dezakegu esaldi

sinpleen kasuan, hau da, esaldian puntuazio-markarik gabe honelako elementuen sekuentzia agertzen bada:

- Aditza
- Kasu nuklearrak (ergatibo, absolutibo eta datiboa)
- Adjuntuak
- Adberbioak
- Nominalizazioak
- Erlatibozko menpeko perpausak
- Konpletibo menpeko perpausak
- Moduzko menpeko perpausak
- Denborazko menpeko perpausak
- Zehargalderak

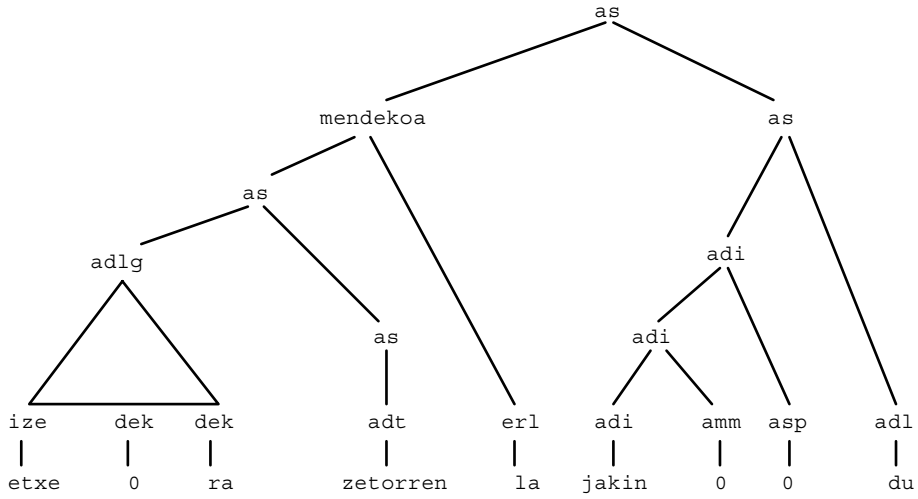
Adibide erreal gisa, *Jakina da gaurko gizonak daraman bizimoduak ez duela antzarik antzinako gizonek zeramatzatenekin* esaldia –euskara-ikasle batek idatzitako corpus batetik atera duguna– osorik analizatzen du gramatikak[17].

7. irudian *'gure etxe polit hark'* izen-sintagmari dagokion analisia ikus daiteke. Hor *'gure'* izenlaguna eta *'etxe polit hura' is2* bilduz *is3* motako osagai bat lortu da, eta berau *'ak'* deklinabide-atzizkiarekin batuz lortu du gramatikak hitz-kate osoa *isk* gisa ezagutzea. 8. irudian *'etxera zetorrela jakin du'* aditz-sintagmari dagokion analisia ikus daiteke. *'etxera zetorrela'* mendeko perpaus konpletibo gisa lotzen zaio esaldi nagusiari.



7. irudia. *'gure etxe polit hark'* izen-sintagmaren zuhaitz sintaktikoa

17 Analizatzaileak onartu du '*antzarik*' hitz okerra. Hori gertatu da hitzak aztertzean morfologiako errore tipikoak tratatzeko gai delako.



8. irudia. *'etxera zetorrela jakin du'* esaldiaren zuhaitz sintaktikoa

4 Semantika

4.1 Sarrera

Semantikaren tratamendua lengoaia naturalaren prozesamenduan diharduten guztiek nahi duten ideala da. Horren arrazoi nagusia da, kasu askotan, hizkuntzaren tratamendu informatikoak adierazpen linguistikoen esanahia atzematea eskatzen duela. Itzulpen automatikoak zein informazio-erauzketarako sistema automatikoek emaitza hobeak dituzte esaldien esanahia kontuan hartuz gero.

Bestalde, semantikaren tratamendua LNPko alderik zailena da, ez baitago inplementa daitekeen esanahiaren teoria orokor bat. Hasieran, ordea, semantika konputazionalean egin ziren lehenengo saioetan, ez zen oinarri teoriko sendorik erabili. Alde horretatik, seguru asko, Montague-ren gramatika izan da teoria semantiko inportanteena (1973). Teoria horren konposagarritasun-printzipioaren arabera posible da eraikitzea osagai sintaktiko baten errepresentazio semantikoa bere azpiosagaien errepresentazioak bilduz. Oso urrun gaude esanahiaren teoria orokor eta oso bat baliatu ahal izateko.

Horren aurrean, analisi konputazionalean tratamendu praktiko bat nagusitzen da. Hau da, ahal dena egitea eskura dauden baliabideekin jokatuz. Hala ere, beti ez da horrelako planteamendu errealistarik izan. LNPren historian zehar, memento batzuetan pentsatu izan zen analisi semantikoa problemarik gabe egin zitekeela. 60ko hamarkada izan zen horietako une bat, adimen artifizialaren memento ospetsuenetan. Orduan pentsatu zen testuen esanahia analiza zitekeela ezagutza linguistikorik gabe. Ondoren, 70eko hamarkadan, batez ere itzulpen automatikora zuzendutako prozesamendu-sistema handiak eraiki ziren. Sistema horietan errepresentazio semantikoa errepresentazio sintaktikoaren ondoren egin beharreko lana izango zen. Gaur egun, aldiz, analisi semantikoaren zailtasunez jabetu dira LNPko arlokoak, eta badakite aukerak mugatuak direla.

Esaldi bat modu automatikoan prozesatu nahi dugunean, tratamendu semantikoaren helburua esaldiaren esanahia lortzea da, hau da, bere edukiaren errepresentazio kontzeptuala sortzea. Horretan, esaldiaren esanahia egitura formal baten bidez adierazi beharko da, eta horrelako adierazpideei *esanahi-adierazpide* deituko diegu. Ezagutza adierazteko moduak fonologian, morfologian, eta sintaxian zuen garrantzizko papera bera du hemen ere, tratamendu semantikoan. Semantikaren kasu honetan adierazpideak ezagutza linguistikoen eta ez-linguistikoen arteko zubi-lana egin behar du eguneroko jardueran hizkuntzaren bidez erabiltzen duguna errepresentatzeko.

Hala ere, *semantika konputazionala* izeneko atal honetan analisi semantikoaz hitz egingo dugunean, era mugatuago batean erabiliko dugu, alegia, testuingurutik independente den esanahi-adierazpideaz hitz egingo dugu. Baina, posible al da hitz egitea *esaldiaren esanahia*ri buruz bere testuingurua kontuan hartu gabe? Ematen du badirela esanahi-aspektu batzuk testuingurutik independenteak direnak, hala nola, hitzen adiera desberdinak bereiztea (bai objektuetan, bai ekintza edo gertaeretan ere), edo nolakoa den esaldi barruko hitzen arteko eragina beren esanahiak elkarri murrizteko. Hemen, horretaz mintzatuko

gara, hau da, testuingurutik (aurreko esaldiak edo hizketa-gaia) independente den esaldien esanahiaz arituko gara. Hurbilpen teorikoak behar dira helburu hori ganoraz lantzeko.

4.2 Oinarrizko kontzeptuak

4.2.1 Interpretazio semantikoaren espezifikazioa. Forma logikoa

4.2.1.1 Semantika eta forma logikoa

Klasikoki esaldi bat ulertzeko egin behar den prozesaketa hiru fasetan banatu ohi da: 1) *analisi sintaktikoa*, 2) *interpretazio semantikoa* eta 3) *interpretazio pragmatikoa.* Azken fase horri *testuinguru eta munduari buruzko interpretazioa* ere deitu izan zaio. Ikuspuntu horretatik interpretazio semantikoaren espezifikazioa honela egiten da:

Sarrera:

Esaldi baten egitura sintaktikoa (lehenengo fasean lortu dena)

Irteera:

Esaldiaren esanahiaren errepresentazioa (forma logiko baten bitartez), baina oraindik testuinguruko aspektuak kontuan hartu gabe. Baztertu egin dira esanahi ulergarria ez duten egitura sintaktiko zuzenak.

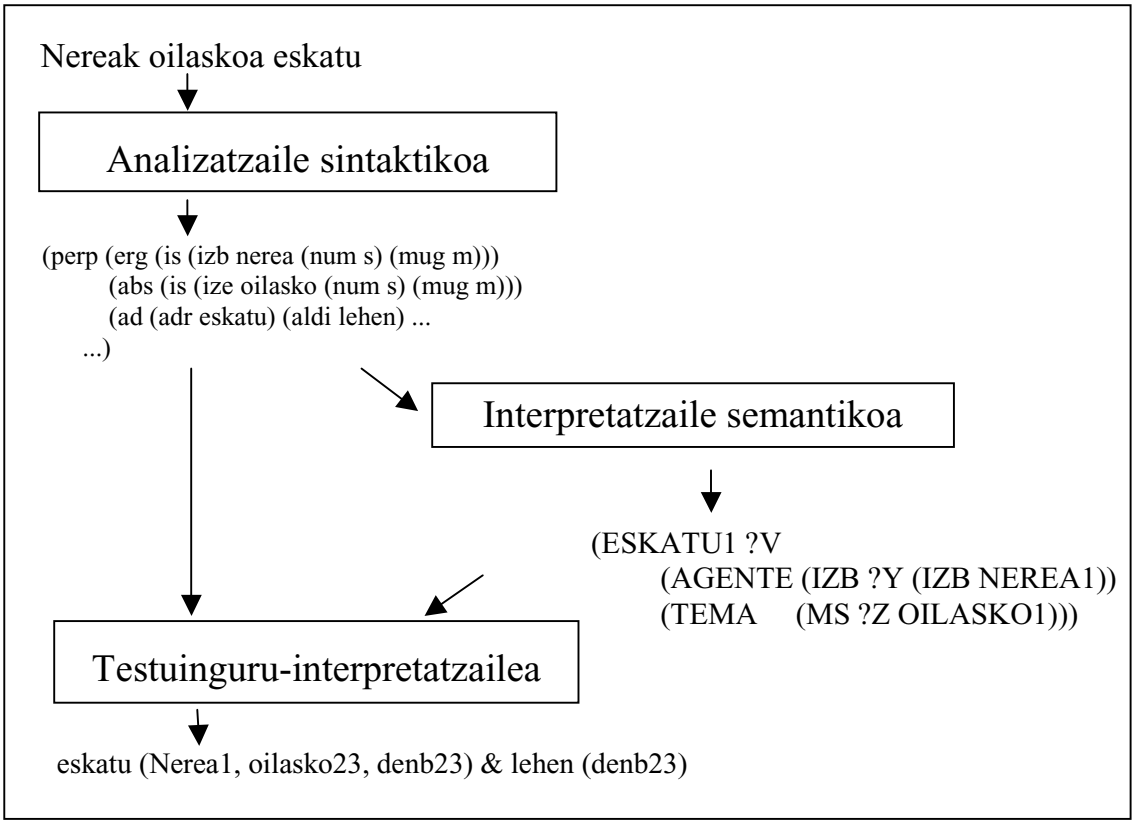
Adibidez: `*Asmo berde kolorebakoak bortizki lo egiten zuten.` Eta hitzen adiera ezegokiak ere baztertu dira.

Adibidez:
```
*Katua(tresna)lotan zegoen
 Katua(animalia)lotan zegoen
```

Bi kontzeptu dira funtsezkoak zehaztapen horretan: hitzen adierak bereiztea (bai objektuetan, bai ekintza edo gertaeretan ere), eta zehaztea noraino diren bateragarriak esaldi bereko hitzen adierak, adiera-konbinazio desegokiak baztertzeko asmoz.

Horrela `*Asmo berde kolorebakoak bortizki lo egiten zuten` esaldiko hitzen adierei erreparatuta, ezin da onartu “asmoek lo egiten dutela”, “bortizki lo egitea” edo “kolorebakoa eta berdea aldi berean izatea”. Esaldi horrek ez du ez hankarik eta ez bururik. Beste alde batetik, “katua” hitza anbiguoa da bi adiera dituelako, baina `*Katua(tresna)lotan zegoen` esaldi horretan ezin da onartu “gurpilak konpontzeko tresnak lo egiten duela”! Badakigu jakin lo animaliek bakarrik egiten dutela eta tresnak ez direla animaliak. Erabaki hauek hartzeko behar ditugun informazio semantikoak lexikotik lor ditzakegu, beti ere aurreko esaldietan esan denari erreparatu beharrik gabe.

Ikus dezagun, esaldi adibide batekin, zer lortzen den analisiaren hiru fase horietako bakoitzean. Hau da esaldia: `Nereak oilaskoa eskatu du`

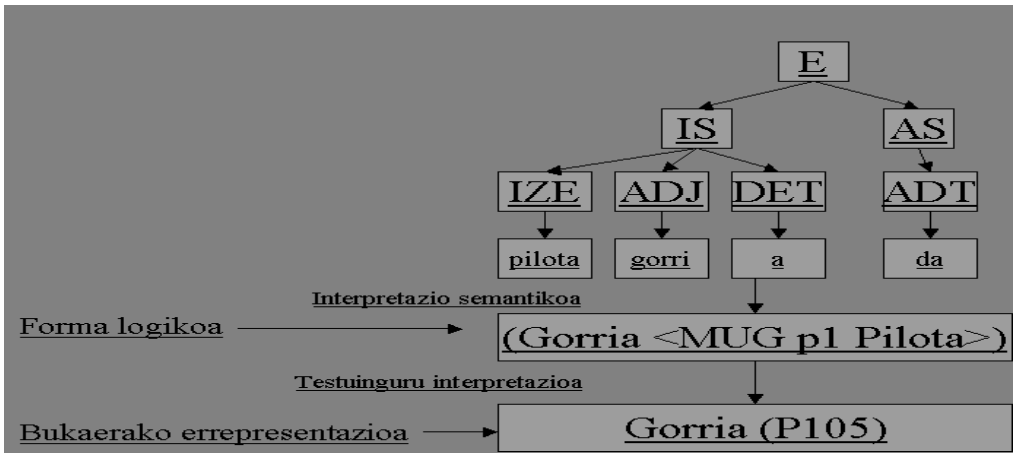


9. irudia. Hiru faseak perpaus baten analisian

Analisi sintaktikoaren emaitzan osagai sintaktikoak bildu dira perpaus-egitura batean. Interpretazio semantikoaren emaitzan, batetik, identifikatu da zein ekintza mota den (ondo definituta egongo den ESKATU1 adiera formala aukeratu da, ESKATU2, ESKATU3... adieretatik bereizten dena). Sintaxi mailako `Nerea` izen berezia eta ergatiboa zena, *agente* kasu semantikopean azaldu da; `oilaskoa` izen-sintagma zena *tema* kasu semantiko gisa azaldu eta OILASKO1 adiera ez-anbiguoarekin lotu da.

Geroago, azken fasean, analisi kontzeptualean, identifikadore semantiko horiek mundu errealeko objektuen identifikadoreekin lotu dira (oilasko23, agian aurreko esaldietako batean aipatua izan dena), eta ekintza semantikoari dagokion asertzio logikoa lortu da gero (`eskatu (Nerea1, oilasko23, denb23)`).

Ondoko beste adibidean azken esanahia lortzeko bi pausoak bereiziko ditugu berriz ere.



Beraz, espezifikazio horren arabera konputazionalki bi prozesu bereizten dira esaldi baten azken esanahia lortzeko:

- **Interpretazio semantikoa.** Esaldiaren esanahi abstraktua lortzen da testuingurua kontuan hartu gabe. Forma logiko baten bitartez adierazten da esaldiaren esanahia. Lexikoetan, hiztegietan, esanahiaz dagoen informazio "estatikoa" erabiliko dugu hemen. Sistematikoki ondo bereizita eta ezarrita dauden *hitz-adierak* dira semantikako atal honen funtsezko osagaia. Sintaxi mailan hitzak (lexiko-sarrera) duen garrantzia, hemen semantikan, adierari dagokio. Osagai sintaktiko baten errepresentazio semantikoa bere azpiosagaien errepresentazioak bilduz lortzen da.
- **Testuinguru eta munduari buruzko interpretazioa** (ulertze kontzeptuala, interpretazio pragmatikoa edo interpretazio kontzeptuala ere deitua izan dena). Esaldiaren esanahi konkretua lortzen da munduari buruzko informazioa eta aurreko esaldien osagaien erreferentzia kontuan hartuta. Forma logikotik esanahiaren azken adierazpidera pasatuko gara.

Esaldiaren esanahia egitura formal baten bidez adierazi beharko da, eta horrelako adierazpideei *esanahi-adierazpide* deituko diegu. Ezagutza adierazteko modu bat baino gehiago aukera daiteke. Adibidez, honako esaldi hau hartuta:

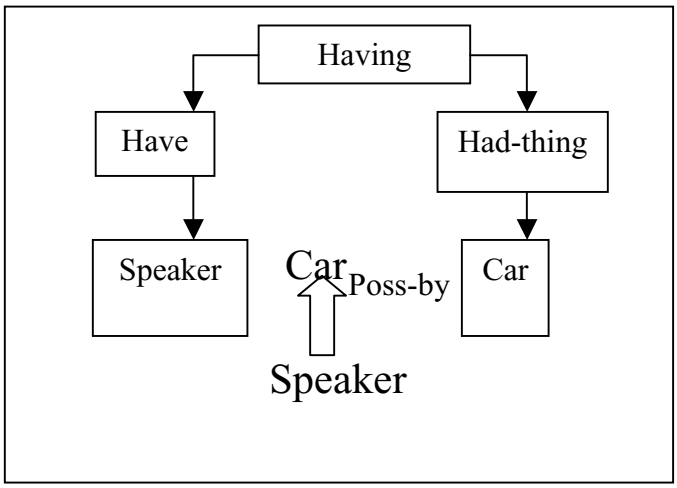
"I have a car"

Esaldi horren esanahia ondoko adierazpide mota hauetan aurkez daiteke:

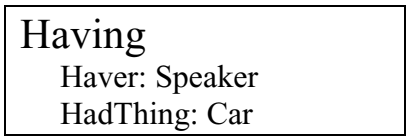
- Forma logikoa**:**

```
badago x, y ((Having x)
              and (Haver speaker x)
              and (Car y)
              and (HadThing y x))
```

- Grafoa:



- Erregistroa**:**

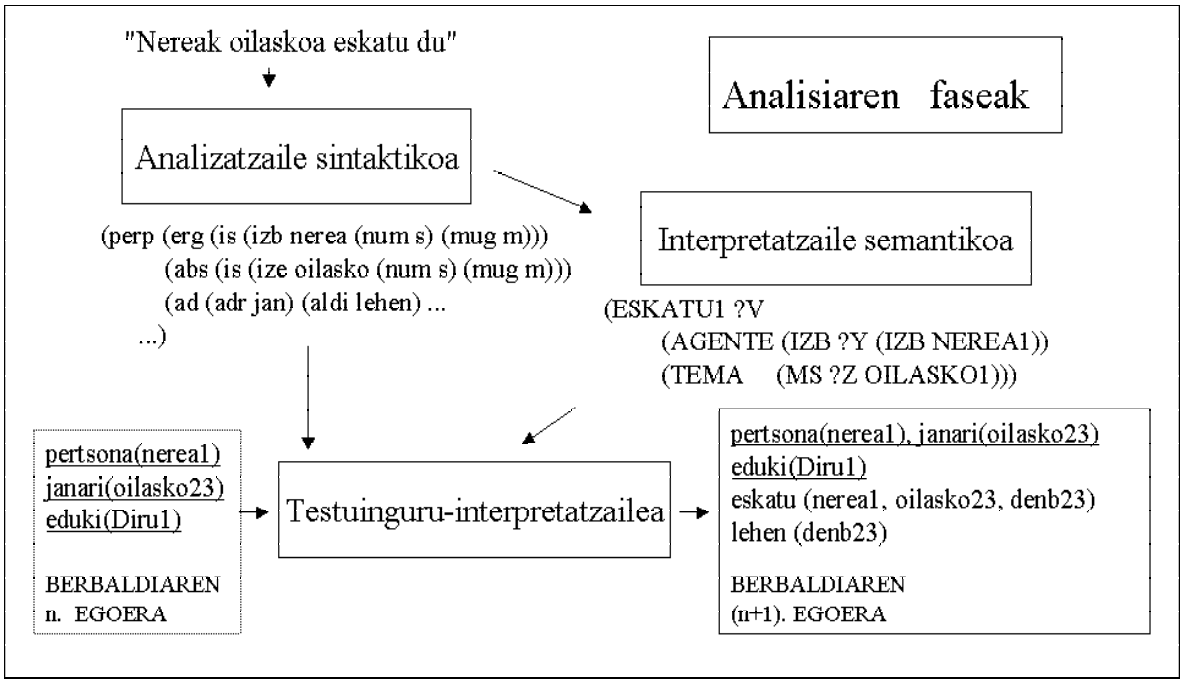


Aurrerantzean, txosten honetan, esaldiaren egitura sintaktikotik sortzen den esanahi-errepresentazioa forma logikoaren bitartez adieraziko dugu beti. Aukera hori hartuta, matematikan oso oinarri sendoa duen predikatu-kalkulu sendoa oinarri-oinarrizko tresnatzat hartzen dugu. Hurbilpen askotan predikatu-kalkuluaren forma logiko hutsa da esaldiaren esanahiaren errepresentazioa, baina hainbat esaldi arrunt adierazi ahal izateko, abiapuntu hori aberastu behar izaten dugu ezaugarri berriekin:

- Testuinguruaren arabera interpretatuko diren terminoen erabilera onartu behar du (izenordainak, determinatzaileak, izen-sintagma mugatuak).
- Kuantifikatzaileak, aditzen denbora eta modua adierazteko ahalmena izan behar du (kuantifikatzaileek, normalean, testuinguruaren arabera zehaztuko dute beren esparrua geroko fasean).

Hurbilpen batzuetan forma logikoak berak aldi berean bi gauza adierazten ditu: esaldiaren esanahi zehatza eta errepresentaziorako lengoaia. Dena dela, forma logikoarekin soilik ez dugu lortzen esanahiaren errepresentazioa, eta hurbilpen horiek, konputazionalki, arazo asko izaten dute gero. Aurreko hamarkadan sortu zen hurbilpen berri batean *egoera* kontzeptu berria erabiliz arazo hori gainditzen zen.

- Hurbilpen horren arabera, *egoera* da munduko hainbat zirkunstantzia biltzen dituen multzo konkretu bat. Forma logikoak egoera batetik beste batera pasatzeko balio digu. Egoera berrian, lehengo zirkunstantzien multzoari forma logikoak sortu dituenak gehituko zaizkio.



4.2.1.2 Esanahiak eta anbiguotasuna

Teoria semantiko bat garatzeko egitura-eredu bat behar da sintaxian egin den bezala. Tratamendu sintaktikoan egitura guztien oinarrizko unitate bat definitu zen (hitza edo morfema) eta horren inguruan egitura konplexuagoak nola eraiki behar diren arautu zen gero. Tratamendu semantikoan ere gauza bera egin beharko dugu: lehenengo **oinarrizko unitatea** aukeratu (hitza, morfema edo esanahiak?) eta gero arautu egin beharko dugu nola sortu egitura semantiko konplexuagoak (esaldi-esanahiak, adibidez). Logikoena unitatetzat esanahia aukeratzea da, eta esanahi gisa hitzen adierak hartzea.

Anbiguotasuna interpretazio semantikoaren prozesuan arazo inportantea izaten da, eta lotuta dago esanahiei. Hiru anbiguotasun mota bereiziko ditugu adierei eta interpretazio semantikoari lotuta:

- Anbiguotasun lexikala: polisemia, homonimia.
- Polisemia: hitz batek erlazionatuta dauden esanahi bat baino gehiago ditu.
- Homonimia: hitz batek beren artean harremanik ez duten esanahi bat baino gehiago ditu.

  Adibideak:

  Nik banku bat dut (eserleku/eraikin) eta Jonek bi. (anb.)
  Nik bi zaldi ditut eta Jonek bat. (Pottoka/….) (lausoa)
  Artoa heldu da (iritsi/umotu).(anb)

- Kategorian oinarritutako anbiguotasuna (inportantea sintaxi mailan).
- Egitura-anbiguotasuna.

  Adibideak:

  Katu eta txakur politak ditut.

  Bi egitura sintaktiko posible:

Katua eta (txakur polita) ditut edo (Katu polita) eta (txakur polita) ditut.

`Ume guztiek txakurra maite dute`

Egitura sintaktiko bakarra, baina bi interpretazio posible: *Nork berea* eta *Guztiek txakur bera*

`Guztiok istripua ikusi genuen`

Lausoa (*vagueness*). Ez dakigu nork edo zenbat lagunek ikusi zuen istripua

Hitzen arteko kokakidetzaren azterketa oso baliagarria da anbiguotasunaren ebazpenean. Adibidez, *Heldu* aditzak hiru adiera ditu euskaraz: *Heldu* (iritsi, ailegatu), *Heldu* (umotu, ondu, zoritu), *Heldu* (eutsi). Hiru adiera horien artean desanbiguatzeko zenbait bide:

- Egitura sintaktikoari kasu eginez desanbigua daiteke batzuetan.
- Hitzen adieren kokakidetzari kasu eginez ere desanbigua daiteke beste batzuetan.

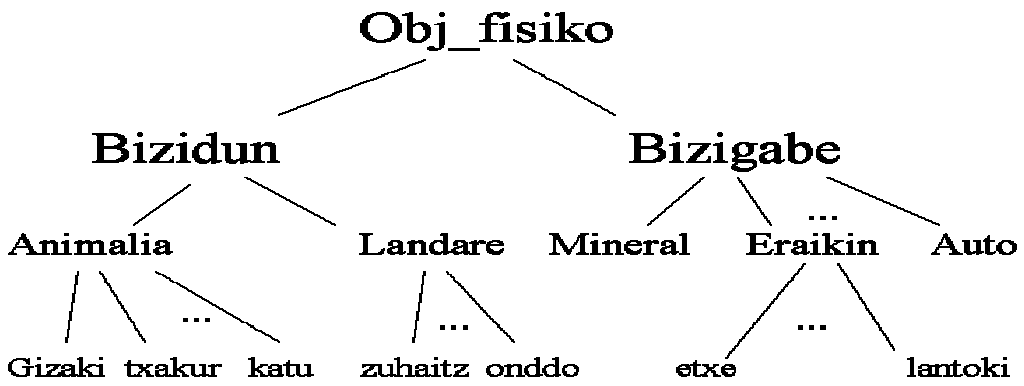
4.2.1.3 Esanahien sailkapena. Ontologia

Hitz asko dago hizkuntza batean eta hitz bakoitzak esanahi bat baino gehiago ditu. Arazoak izan ditzakegu horrenbeste esanahi errepresentatzeko. Esanahi horiek elkarren artean erlazio handiak dituztenez, erlazio horiek erabil ditzakegu errepresentazio trinkoago bat lortzeko. Esanahiak klase/azpiklase erlazioaren arabera hierarkikoki antola daitezke, ontologia bat sortuz.

> *"Ontologies are agreements about shared conceptualizations. Shared conceptualizations include conceptual frameworks for modeling domain knowledge; content-specific protocols for communication among inter-operating agents; and agreements about the representation of particular domain theories…" "Ontologies: principles, methods and applications*". Mike Uschold and Michel Gruninger. *The Knowledge Engineering Review,* Vol. 11:2, 1996, 93-136.

Kontzeptuen sailkapena betidanik izan da interes handiko gaia. Aristotelesek berak ere klase batzuk definitu zituen: objektu fisikoak, kantitateak, kalitateak, erlazioak, lekuak, denbora, posizioa, egoera, ekintzak eta afekzioak. Horiei gehi dakizkieke beste batzuk: gertaerak, planak, kontzeptuak eta ideiak.

Non klase garrantzitsuenak *gertaerak, ekintzak* eta *egoerak* diren.



10. irudia. Ontologiaren adibidea

4.2.1.4 Forma logikoaren oinarrizko lengoaia. Forma sasilogikoak

Lehenengo ordenako predikatuen logika izeneko formalismo matematikoa erabiliko dugu hemen. Formalismo honen oinarrizko kontzeptuak **atomoak** edo **konstanteak** dira: ontologia batean antolatu eta anbiguotasunik gabe bereizi diren esanahiak. Atomoen artean terminoak eta predikatuak honela bereiziko ditugu:

- **Terminoak:** munduko objektuak (objektu fisikoak, objektu abstraktuak, egoerak, ekintzak eta gertaerak barne).
- **Predikatuak:** propietateak eta erlazioak zehazteko definituko dira, beti ere argumentu batzuekin erabili beharko direnak. **Proposizio** bat osatuko da predikatu batekin eta horrek behar dituen argumentuekin, terminoen bidez adierazita, noski.

Hala ere, oinarrizko forma logiko horietan *lehenengo ordenako predikatuen logika* bere horretan bakarrik erabiliz, hizkuntzaren azpimultzo oso mugatu bat baino ezin da adierazi. Beste baliabide batzuk gehitu behar zaizkio formalismo horri, testu errealetako esaldien esanahia adierazi ahal izateko. Horien artean aipatu behar dira hauek:

- Kuantifikadoreak
- Adjektiboak
- Pluraltasuna
- Eragile modalak (nahi, uste...), aldia (orain, gero...), eta abar

Forma logikoen oinarrizko adierazpidea ikusita, aditzak eta egoerak forma logikoan adierazteko aukerei erreparatuko diegu. Aditzen eta beren argumentuen arteko harremanei kasu eginez, erlazio sintaktikotik harantza joanez, horrelako esaldien antzekotasuna forma logikoan islatu ahal izateko erlazio semantiko abstraktuen multzo bat definitu izan da (rol tematikoak).

Rol tematikoak (Kasu semantikoak ere deituak) hauexek izaten dira:

Rolak eta azpirolak		**Definizioak**
AGENTEA	AGENT	Gertaera sortu duen nahitako kausalitatea
INSTRUMENTOA	INSTRUMENT	Gertaera sortzen duen instrumentua/indarra
TEMA/PAZIENTEA	THEME	Gertaerak eragiten duen objektua
ESPERIMENTATZAILEA	EXPERIENCER	Gertaerari psikologikoki edo fisikoki lotuta dagoen pertsona
ONURADUNA	BENEFICIARY	Norentzat egin den ekintza
NON/KOKAPENA	LOCATION /AT-LOC	Uneko kokapena
NOREN/EDUKITZAILEA	POSSESSOR /AT-POSS	Uneko edukitzailea
ZER-BALIOA	AT-VALUE	Uneko balioa
NOIZ	AT-TIME	Uneko denbora
NORA/HELBURUA	TO-LOC /DESTINATION	Bukaerako kokapena
NORENTZAT/HARTZAILEA	TO-POSS /RECIPIENT	Bukaerako edukitzailea

NORA-BALIOA	TO-VALUE	Bukaerako balioa
NONDIK/JATORRIA	FROM-LOC / SOURCE	Hasierako kokapena
NORENGANDIK	FROM-POSS	Hasierako edukitzailea
NONDIK-BALIOA	FROM-VALUE	Hasierako balioa
AGENTEKIDEA	CO-AGENT	Ekintzaren bigarren agentea
TEMAKIDEA	CO-THEME	Ekintzaren bigarren tema
BIDEA	PATH	Zein bidetatik zerbait pasatzen den

Aditzak sailka daitezke exijitzen duten rol tematikoen arabera. Dena dela, komeni da bereiztea rolen artean aditzarekin harreman estua dutenak, eta harreman esturik ez dutenak. Hain zuzen ere, aditzen osagarrietan sorburua dutenei buruz, aditzekin barne-harremana estuagoa dutela esan daiteke. Adibidez, *joan* aditzak azpikategorizatzen du *nora* kasu sintaktikoa eta kasu horrek NORA/HELBURUA kasu semantikoa adierazten du. Rol mota hauek inportanteak dira aditzen azpikategorizazioan; normalean *barne-rolak* edo *argumentuak* deitzen zaie, beste kasuei *adjuntu* esaten zaie. Hau esan daiteke: "aditz bat erabiltzen den guztietan rol bat beharrezkoa bada, orduan rol hori argumentua izango da" baina hau ez da hain erraza; izan ere, badira barneko rolak aukerakoak direnak.

4.2.1.5 Konposagarritasuna

Interpretazio semantikoa prozesu konposizionala da. Hau da, elementu baten esanahia bere osagaien esanahiak konbinatuz lortzen da. Konposizioan oinarritzen diren teoriak interesgarriak dira, esaldiaren esanahiaren errepresentazioa inkrementalki sortzen delako. Testuingururik gabeko sintaxirako gramatika-eredua konposizionala da. Erregelak azpiosagaien kategoriaren arabera aplikatzen dira eta ez diote begiratzen azpiosagai horien barne-egiturari. S → NP VP erregela aplikatzen denean ez da aztertzen zer nolako NP edo VP dagoen, edozein NP edo VP osagaitarako definitzen da.

Teoria konposizionalen arabera, osagai sintaktiko baten esanahia bere azpiosagai sintaktikoen esanahiak konposatuz lortzen da. Funtzio baten bitartez adierazten da konposizio hori. Eredu hau erabiliz esanahiak lortuko dituen gramatikaren eraikuntza eta mantentzea ikaragarri errazten da. Hala ere, konposagarritasun hutsean datzan teoria ez dabil beti; gehienetan bai, konposagarritasuna aplikagarria da, baina zenbait kasutan ezin da aplikatu:

- Egitura sintaktikoa ez dator bat forma logikoaren egiturarekin.
- Hitz anitzeko terminoak arazo-iturburu izaten dira, horrelakoetan esanahiak ez baitu zerikusirik osagaien esanahiekin.

Lambda kalkulua deituriko formalismoa oso lagungarri gertatzen da eta sarritan ahalbidetzen du osagai sintaktiko partzial baten interpretazio semantikoa definitu ahal izatea.

Lehen ikusi ditugun teknikek ez dituzte konpontzen anbiguotasunaren arazoak. *Zubi* izen arruntak 3 esanahi ditu: a) erreka edo hutsunea gainditzeko egitura, b) asteburu luzea, c) hortzen arteko protesia. Aurretik ikusitako formalismoen arabera, 3 forma logiko sortuko lirateke, bakoitza esanahi batekin eta, gero, hiru forma logiko horien artean erabaki behar da agertzen den esaldiko zer zentzuri dagokion.

Baina esaldiko beste hitzen adieren arabera adiera batzuk baztertu ahal izango dira.

> *Parisen egongo naiz hil-bukaerako zubian*
> *Ibaiak zubia zeharkatzen du.*

Mota semantikoen hierarkiak eta hautapen-murriztapenak lagungarri suertatzen dira anbiguotasuna kentzeko; zentzu posibleen artean, batzuk baztertzen lagun dezakete.

4.3 Hurbilpenak

Semantikaren prozesamenduari ekiterakoan dauden joera nagusiak honako hauek ditugu:

- Errepresentazio sintaktikoen gainean informazio semantikoa gehitzea. Oinarrian tratamendu sintaktiko osoa dagoenean sistemarik erabiliena da hau. Analisi sintaktikoan hitzak aukeratzen laguntzen duen informazio semantikoa esleitzean datza. Nagusiki, esleitzen den informazio semantikoa bi tipotakoa izaten da: izenek ikuspegi semantikotik deskribatzen laguntzen duten ezaugarri semantiko batzuk jasotzen dituzte eta aditzek (eta gainerako predikatuek) beren posizio argumentalak markatuak dituzte ezaugarri semantikoen bidez. Modu horretara, sistema sintaktikoek esaldietako hitzen esanahiaren gaineko nolabaiteko kontrola izan dezakete. Eta, beraz, hitzik egokiena hautatu kasu bakoitzean. Hurbilpen honen arazorik latzena hitzen semantika ezaugarri semantikoen bidez errepresentatzeko modu egokia ez egotean datza. Eginkizun hori ezinezkoa dela dirudi. Dena den, zenbait ezaugarri semantiko sinple (gizaki+/-, zehatza+/-, zenbakarria+/-, etab.) erabilgarri suertatzen dira interpretazio posibleen artean zein den egokia erabaki behar denean.
- Sintaxiaren ondoko maila batean kokatzen den semantika. Hurbilpen honen estrategia maila sintaktikotik eratortzen den egitura-semantikoa maila sintaktikoaren ondoren sortzean datza. Semantikaren tratamendu hau, bereziki itzulpen automatikoaren aldetik sustatu da 70 eta 80ko hamarkadetan. Urte haietan hizkuntzaren prozesamendua gramatika sortzaile/transformatzailearen gertuko ikuspegitik abiatzen zen. Ikuspegi hori analisi semantikoa independenteki eta sintaxitik eratorrita egitean gauzatzen zen. Egun, sistema erreal gutxik jarraitzen diote estrategia horri, errepresentazio semantikoaren maila bera eta sintaxitik semantikarako proiekzioa zailak baitira modu koherente eta oso batez definitzen.
- Errepresentazio sintaktikoan txertatzen den semantika. Gramatikaren baitan, kategoria konplexuetatik eratortzen den ikusmolde berriak modu berri bat ahalbidetzen du semantika hizkuntzaren prozesamenduan bideratzeko. Kontua da, deskribapen semantikoa ez dela sintaxitik aparte egingo, baizik eta, sintaxiak eta morfologiak osatzen duten deskribapen linguistikoaren parte bezala hartuko da. Hartara, semantika errepresentazio sintaktikoarekin integratzen da. Beraz, erlazio sintaktikoak kalkulatzerakoan, semantikoak ere kalkulatu egingo dira, eta bi alderdiak kontuan hartzen dituen errepresentazio bakar bat erdietsiko da. Estrategia hau

darabilten prozesadore erreal gehienetan, errepresentazio semantikoa ezaugarri semantiko estrukturalenetatik abiatzen da: argumentu-harremanak, denbora eta aspektua, determinazioa... Aldiz, ia inoiz ez dira kontuan hartzen semantika lexikalaren atalean lantzen diren fenomenoak, hala nola, sinonimia, antonimia, hiperonimia eta meronimia, etab. Hala ere, semantikaren eta baterakuntza-formalismoen sintaxiaren arteko erabateko integrazioak egingarria dirudi. Oinarriak ezarrita daude eta integratze hori gauzatzea denbora-kontua da.

- Edozein teoria semantikorekiko independentea den markatze semantikoa. Sintaxian zein semantikan tratamendu semantiko oso batek dituen zailtasunak kontuan harturik, zenbait ikertzailek konputazionalki kostu txikiagokoak diren hurbilpenak landu dituzte. Horrela, testuan bertan zuzen-zuzenean markatze semantikoak egiten dituzten sistema batzuk sortu dira. Anotazio semantiko horiek egiteko ez da aurretik maila altuko prozesamendu sintaktikorik behar. Kasu horietan, testuak markatzeko teknikak baliatzen dira, maiz oinarri estatistikokoak (desanbiguatzaile morfologiko batzuetan baliatzen direnen antzekoak). Bestalde, ezaugarri-sistema hauen ezaugarri orokorra da erabiltzen dituzten marka semantikoak oso baldintzatuta daudela tratatzen duten testu motagatik. Hau da, erabiltzen dituzten etiketa semantikoak kanpo semantiko partikular bateko testuentzat izango dira baliagarriak (adibidez, juridikoa). Baina, ezin izango lirateke erabili testu orokorrak edota bestelako kanpo semantiko batzuetakoak tratatzeko. Sistema hauek funtzio espezifiko batzuei erantzuten diete semantikaren tratamendu osoa eta koherentea egin ahal izateko sistemaren baten gabeziaren aurrean. Adibidez, testu zehatz batzuetan adierak desanbiguatzeko balia daitezke. Baina, ezin dira analisi semantikorako sistema gisa hartu, horretara iristeko behar adina orokortasunik ez dutelako. Badirudi prozesamendu- eta errepresentazio-sistema osoagoak eta semantika erabat kontuan hartzen dutenak sortzen diren neurrian, sistema mugatu hauek indarra galtzen joango direla. Azken finean, semantikaren tratamenduari dagokion arloa gehien aurreratu behar dutenetariko bat dugu. Gramatika sintaktikoak ezaugarri semantikoekin osatzeaz gain, gaur egungo lanen etorkizuna aipatu ditugun azken bi hurbilpenen ildotik doa: semantika-kategoria konplexudun gramatiketan integratzea eta teoria linguistikoetatik independente den testuen markatze semantikoa. Lehenengo hurbilpenak tratamendu konputazional orokorretarantz aurrera egiten laguntzen du; eta, bigarrengoak, tratamendu semantiko orokor bat ez dagoen bitartean, desanbiguazio semantikoaren inguruko lanak gauzatzen ditu.

4.3.1 Gramatika-erlazioak

Sistema hauetan analizatzaile sintaktikoak ez ditu ematen egituraren zehaztapen guztiak, semantikarako inportanteak diren ezaugarriak baino ez ditu ematen. Erlazio hauei *gramatika-erlazio* edo *menderakuntza gramatikalak* esaten zaie. Subjektu, objektu logiko, zehar-objektu eta abarren erlazioak erlazio gramatikalen artean kokatzen dira. Erlazio bakoitza errepresentatzeko hirukote-egitura aukeratu da. Formatua <*aldagaia erlazioa balioa*> da.

Interpretatzaile semantikoak esaldiaren interpretazioa lortuko du dependentzia-erlazio hauetan oinarrituz. Analisitik ateratzen den hirukotea erregela bateko ereduarekin bat datorrenean, dagokion forma logikoaren zatia sortzen da, eta bukaeran zati guztiak konbinatzen dira formula logiko bat lortzeko.

4.3.2 Gramatika semantikoak

Interpretazio semantikoa aplikazio konkretu baterako egin nahi bada, badira moduak analisi sintaktiko eta semantikoak eraginkorragoak izan daitezen.

Eraginkortasun hori anbiguotasunaren tratamenduaren sinplifikazioan oinarritzen da. Hitzak testuinguru konkretuetan ager daitezkeen ideiaz baliatzen dira eta, beraz, hitzaren hainbat interpretazio ez dute kontuan hartu behar.

Gure adibidean, bezeroei ematen zaien hegaldiei buruzko informazioarekin du zerikusia. Abioei buruzko informazioa datu-base batean dago eta datu-baseari galderak eginez lortzen da. Datu-basean galdeketetan sortzen den lengoaia aztertu behar da egitura sintaktiko eta esanahi tipikoak lortuz. Egitura sintaktiko batzuk testuinguru semantiko konkretuetan azalduko dira; kasu hauetan, erregela sintaktikoetan aspektu sintaktikoak, analisi prozesuan lagunduko duten ezaugarri semantiko batzuen bitartez ordezka daitezke eta, horrela, sinplifikatuko da prozesu osoa.

Gure adibide honetan, izen-sintagmek honako egitura dute:

The flight to Chicago
The 8 o'clock flight
The first flight out
Flight 457 to Chicago

Izen-sintagma hauek analizatzeko gramatika honako hau izan daiteke:

```
NP  → DET CNP   (the flight)
CNP → N         (flight)
CNP → CNP PP    (flight to Chicago)
CNP → N PART    (flight out)
CNP → PRE-MOD CNP (8 o'clock flight)
NP  → N NUMB    (flight 457)
```

Erregela hauekin forma ez-zuzenak ere lor daitezke, adibidez:

**the city to Chicago*
**the 8 o'clock city*
**the first city out*
**city 567*

Hori saihesteko informazio semantikoa erants dakieke erregelei. Erregela hauetan propietate semantikoa duten kategoria lexiko berriak azalduko dira. Adibidez FLIGHT-N (*hegaldi* esanahiarekin zerikusia duten izenak). Hau dena kontuan hartuta aurreko erregelak berridatz daitezke:

```
FLIGHT-NP    → DET FLIGHT-CNP         (the flight)
FLIGHT-CNP → FLIGHT-N                 (flight)
FLIGHT-CNP → FLIGHT-CNP FLIGHT-DEST   (flight to Chicago)
FLIGHT-CNP → FLIGHT-CNP FLIGHT-SOURCE(flight from Chicago)
FLIGHT-CNP → FLIGHT-N FLIGHT-PART     (flight out)
```

```
FLIGHT-CNP → FLIGHT-PRE-MOD FLIGHT-CNP(8 o'clock flight)
FLIGHT-NP   → FLIGHT-N NUMB          (flight 457)
CITY-NP       → CITY-NAME            (Chicago)
CITY-NP      → DET CITY-CNP          (the city)
CITY-CNP    → CITY-N                 (city)
CITY-CNP→ CITY-MOD CITY-CNP CITY-MOD-ARG(nearest city to Dallas)
```

Faltako lirateke beste erregela batzuk:

```
FLIGHT-DEST → to CITY-NP
FLIGHT-DEST → for CITY-NP
```

Goi-mailako egitura sintaktikoak ere egongo dira:

```
TIME-QUERY → When does FLIGHT-NP FLIGHT-VP ?
```

Laburbilduz, kategoria sintaktiko eta semantikoen arabera adierazten diren gramatikei gramatika semantikoak deritze. Ez dago oso garbi non dagoen gramatika sintaktikoen eta semantikoen arteko muga. Normalean, gramatika semantikoak erregela askoko gramatikak dira baina erregelak eraikitzeko prozesua motzagoa da. Aplikazio konkretuetarako egokiak, baina domeinua aldatuz gero, gramatika osoa berreraiki behar da.

4.3.3 Patroi-parekatzea

Helburu mugatuko hainbat domeinutan, probetxagarria gerta daiteke domeinuaren egitura tipikoak baliatzea interpretazio semantikoaren prozesuan. Egunkarietan azaltzen diren negozio-eragiketei buruzko laburpenak aztertuz gero, horri buruz ematen diren egunkarietako berri guztiek eskema finko bati jarraitzen diotela ikus daiteke: beti azalduko da erosten duena, erosi dena, zer preziotan erosi den, nori erosi dioten, eta abar.

Eredu sinple batzuk definitzea da teknika honen gakoa. Eredu sinple horiek domeinuko informazio zatiak adieraziko dituzte. Informazio zati horien bitartez osatuko da tarea errepresentatuko duen eskema orokorra.

Hego Amerikako eraso terroristei buruzko domeinuaren inguruan laburpenak egitea bada aplikazioko tarea ondoan azaltzen den eredua izan daiteke egokia.

TERRORIST INCIDENT	
DATE	date
LOCATION	city/state/country
TYPE	e.g. bombing
STAGE of EXECUTION	e.g. accomplished, planned
INSTRUMENT	e.g. bomb, gun
PERPRETATOR NAME	e.g. FMLN
PHYSICAL TARGET	e.g. car, house
HUMAN TARGET	e.g. president
NATIONALITY TARGET	e.g. San Salvador
EFFECT	e.g. no injury

Ideia nagusia hau da: sarrerako testuan ereduak definitu, non eskema orokorreko atributuak identifikatuko diren.

Adibidez: *take* aditza, gizakia den zerbait deskribatzen duen sintagma, eta *hostage* hitza agertzeak hitz-sekuentzia batean, TERRORIST-INCIDENT eskemako HUMAN-TARGET atributuaren balioa adierazten du. Hori guztia eredu honen bitartez adierazten da:

```
take  <HUMAN>  hostage   →
     (TERRORIST-INCIDENT  HUMAN-TARGET  1)
```

lan horretan analizatzaile partzialak egokiak dira. Analizatzaile orokorrak oso garestiak baitira eta gainera ez da oso errealista edozein esaldi onartuko duen analizatzailea egin daitekeela pentsatzea. Analizatzaile partzialak "puskak" hartuko ditu: izen-sintagmak eta aditz-sintagmak; analizatzaileak preposizioak eta loturazko partikulak ere bereiz ditzake. Horretaz gain, zatien mota semantikoaren informazioa behar du; normalean zatiaren gunetik hartuko dira.

Ad. : Guerrillas attacked Merino's home in San Salvador five days ago with explosives.

Erabilitako lexikoa:

Ago	(AGO)
Attacked	(V VFORM past TYPE attack)
Days	(DATEUNIT)
Explosives	(N TYPE WEAPON)
Five	(NUMB)
Guerrillas	(N TYPE HUMAN-GROUP)
Home	(N TYPE LOC)
In	(P TYPE IN)
Merino	(NAME TYPE PERSON)
San-Salvador	(NAME TYPE LOC)
with	(P TYPE WITH)

Analizatzaileak lortuko dituen zatiak:

```
(NG Guerrillas TYPE HUMAN-GROUP)
(VG attacked TYPE ATTACK VFORM PAST)
(NG Merino's home TYPE LOC)
(P in TYPE in)
(NG San-Salvador TYPE LOC)
(NG (BEFORE-NOW 5 days) TYPE DATEUNIT)
(P with TYPE WITH)
(NG explosives (TYPE WEAPON))
```

Eredu sinple batzuk:

```
P1     <HUMAN> <ATTACK> <LOC>      → (INCIDENT ATTACK
                                       PERPETRATOR NAME 1
                                       PHYSICAL TARGET 2)
P2     <IN> <LOC>                  → (LOCATION 2)
P3    <DATE>                       → (DATE 1)
P4    <WITH> <WEAPON>              → (INSTRUMENT 2)
```

Lortzen den analisia:

```
(INCIDENT ATTACK
      PERPETRATOR NAME Guerrillas
      PHYSICAL TARGET Merino's home)
(LOCATION San Salvador)
 (DATE five days ago)
(INSTRUMENT explosives)
```

Hau adibide bat besterik ez da izan. Gauzak konplexuagoak egiten diren heinean, eskema orokorrak eta ereduak konplexu bihurtuko dira.

5 Pragmatika

Perpausen arteko eraginak ugariak eta konplexuak dira; izan ere, gehienetan, perpaus bat ezin da osorik ulertu aurreko perpausak kontuan hartu gabe. Ikus dezagun hori zenbait adibideren eskutik. Adibide bakoitzean perpaus pare bat aurkeztuko dugu eta ondoren galdera bat. Ez du ezelango arazorik izango inork galdera horri erantzuten, baina konputagailu batek erantzutea nahiko bagenu... A zelako komeriak! Zein erraza den pertsona batentzat eta zein zaila konputagailuarentzat! Izan ere, adibide hauek azalduz bi helburu lortu nahi ditugu: batetik, agerian uztea perpaus bat ondo ulertzeko aurreko perpausak ere ulertu behar direla, eta bestetik, horrelako lanak egin ahal izateko konputagailuak beharko lituzkeen tresnak eta ezagutzak identifikatzea

1. **Perpausak:** *Ikerrek txerriak ekarri zituen. Zikin-zikin geratu zen.*

   Galdera: Zein geratu zen zikin?

   Pertsona batek arazorik gabe *Iker* erantzungo du. Baina, hori automatikoki ulertzeko, konputagailuak jakin behar ditu hainbat gauza: 1) *Zikin-zikin geratu zen* esaldian subjekturik ez dagoenez, aurreko esaldi batean aipatu den norbait edo zerbait izan daitekeela. 2) Horrela balitz, *geratu zen* aditzaren subjektu eliptikoak bat etorri behar duela aditzarekin numeroan (singularra), *3) txerriak* plurala dela, 4) *Iker* singularra, 5) eta beraz, *Zikin-zikin geratu zen* hori aurreko perpausean singularra den *Iker* horri dagokiola. Hau guztia ikusita, argi dago diskurtsoari buruzko ezagutza behar dela bigarren esaldia ondo ulertzeko, alegia, eliptiko dagoen objektu bat testuinguruan nola bilatu behar den jakiteko.

2. **Perpausak:** *Jonek baloi zuria nahi zuen. Anek ere hura nahi zuen.*

   Galdera: Zeren aipamena da “hura” izenorde hori? Ane adierazi nahi du? Baloia? Jon? Besterik?

   Pertsona batek arazorik gabe ulertuko du “baloia” adierazi nahi duela, hain zuzen, Jonek nahi zuen baloi zuri bera. Baina hori ulertzeko konputagailuak jakin behar du “ere” partikula dagoenez, aurreko esaldiaren antzekotasuna ematen zaigula, esangura antzekoa izango dela, *Jon*-en ordez *Ane* jarrita, eta beraz, *hura* izenordea lehengo baloiaren erreferentzia dela. Adibide honetan, beraz, diskurtsoari buruzko ezagutza bakarrik behar da ondo ulertzeko.

3. **Perpausak:** *Anek liburu bat erosi zuen. Apurtuta zegoen 11. orria.*

   Galdera: Zein da *“11. orri”* hori? Zuhaitz bateko orria da? Liburu batekoa? Edo Internetekoa?

   Pertsona batek arazorik gabe ulertuko du Anek erosi berri zuen liburuko 11. orria dela. Baina hori konputagailuak ulertu ahal izateko behar du jakin liburuak orriak dituela, 11 orri gutxienez gainera, eta 11. orri hori aurreko esaldiko liburukoa dela testuinguruan beste orri posiblerik ez dagoelako (zuhaitzik, Interneteko helbiderik, beste libururik...). Adibide honetan, beraz, munduari buruzko ezagutza (liburuek orriak dituzte) ere behar da ondo ulertzeko.

Galdera: Zer gertatu da lehenago "*Anek liburua erostea"* edo "*orria apurtzea"*?

Pertsona batek arazorik gabe erantzungo du *orria apurtzea.* Baina konputagailuak ondo jakin behar du perpausetako aditzen aspektua eta aldia aztertzen, *apurtuta zegoen* azaltzen denean lehenago gertatua izan zela bereizteko.

4. **Perpausak:** *Patxi Parisera joan zen lana zela eta. Hendaian hartu zuen gau-trena.*

Galdera: Nora zihoan Patxik Hendaian hartu zuen trena?

Pertsona batek arazorik gabe ulertuko du Parisera zihoala Patxik hartu zuen trena. Baina hori konputagailuak ulertu ahal izateko, batetik, lehenengo perpausean harrapatu behar da Patxik Parisera joateko helburua zuela; beste alde batetik, jakin behar da *trena hartzea* tokietara joateko helburuak betetzeko modu bat dela; gainera, ez da guztiz beharrezkoa, baina jakin behar da Hendaiatik Parisera doana ohiko tren-bidaia dela, eta gau-trena ere badela; eta bukatzeko jakin behar da lotzen bi esaldietako esanahiak, alegia, lotzen Hendaiatik aterako den trena Parisera joateko helburua betetzeko aukera ezin hobea dela eta besterik ezean hori dela bi esaldien esanahia lotzeko hipotesirik onena. Adibidea ulertu ahal izateko behar izan dira munduari buruzko ezagutza (badela gau-tren bat Hendaiatik Parisera joaten dena, eta tren bat hartzea aukera bat dela norabait joateko helburua betetzeko), eta diskurtsoari buruzkoa ere (bi perpaus lotuta egon daitezke, batak helburu bat aurkeztu eta besteak helburu hori betetzeko aukera bat ematen badu).

5. **Perpausak:** *Maitek Lizarrara joan nahi zuen arrantzara. Baina kaina apurtuta zeukan.*

Galdera: Aipatutako kaina zer da? Garagardoa, gune hezeetako landarea edo arrantza-tresna?

Ez dago zalantzarik inorentzat: testuinguru horretan *arrantza-tresna* da. Baina konputagailuarentzat zailagoa da, jakin behar da arrantza dela-eta zer erabili ohi den; era berean jakin beharko litzateke zer jardueratan ager daitezkeen *"kaina garagardoa"* eta *"kaina landarea"* (munduari buruzko ezagutza), eta bukatzeko, testuinguru horretan hiru adieren artean zentzuzkoena arrantza-tresna dela ere bereizten jakin behar da (diskurtsoari buruzko ezagutza).

Galdera: *Besterik jakin gabe, uste duzu Maite arrantzara joan zela?*

Ezetz erantzungo luke edonork berehala. Konputagailuak, ordea, jakin beharko luke kaina ezinbesteko tresna dela Lizarrara arrantzara joateko.

Beraz, *arrantzara joatea* bezalako ohiko jarduerei buruz hainbat gauza jakin beharko ditu: parte-hartzaileak, tresnak, martxan jarri ahal izateko aurrebaldintzak, lortzen diren helburuak, zelako ekintza partzialak gauzatzen diren eta zein ordenatan...

Hori ez da batere sinplea! Batez ere edozein testuingurutan, edozein esaldi, edozein jarduerari buruz ulertu nahi badugu.

6. **Perpausak:** *Bart elurra egin zuen mara-mara. Gaur ez dugu eskolarik izango.*

Galdera: *Zergatik ez dugu eskolarik izango?*

Umeek ondo badakite elur asko egin eta biharamunean ez dela izaten eskolarik. Munduari buruzko ezagutza hori ere barruan izan beharko luke konputagailuak, edo bestela dedukzioak egin ahal izateko ahalmen handia beharko luke. Adibidez: elurra mara-mara egiten duenean errepideak egoera txarrean egoten dira, egoera horretan arriskutsua da ibiltzea, horrelakoetan auto ugari ibiltzen bada egoerak txarragora jotzen du, beraz, hobe mugimendu gutxiago bada, beraz, eskolak-eta itxita egoten dira egun horietan.

Maila horretako hainbat eta hainbat gauza dakizkigu. Horiek guztiak konputagailu batean sartzea ez da batere sinplea! Ulermena gai konkretu bati buruz mugatzen badugu, agian zerbait lor liteke konputagailuarekin, baina, oro har, edozeri buruz ulertzea eskatu nahi badiogu... jai dugu!

7. **Perpausak:** *Izaskunek auto berria erosi nahi zuen. Erretzeari utzi zion.*

   Galdera: *Zergatik utzi zion erretzeari?*

   Pertsona batek arazorik gabe ulertuko du auto berria erosi ahal izateko dela. Baina hori ulertu ahal izateko, konputagailuak helburu eta planifikazioari buruz jakin behar da. Batetik, lehenengo perpausean harrapatu behar du Izaskunek auto berria erosteko helburua zuela; beste alde batetik, jakin behar da autoak garestiak direla eta horrelakoetan *dirua aurreztea* helburua betetzeko modu bat dela; dirua aurrezteko modu bat erretzeari uztea izan daitekeela ere jakin behar da; eta bukatzeko jakin behar da lotzen bi esaldietako esanahiak, alegia, lotzen auto berria erosteko helburua betetzeko Izaskunek dirua aurrezteko plan bat martxan jarri duela; tabakoan ez gastatzea dirua aurrezteko plana gauzatzeko era bat dela; eta hori dela bi esaldien esanahia lotzeko hipotesirik onena. Adibidea ulertu ahal izateko behar izan dira munduari buruzko ezagutza (auto berriak garestiak dira, gauza garestiak lortu ahal izateko plan posible bat aurreztea da, erretzeari utzita aurrezten da) eta diskurtsoari buruzkoa ere (bi perpaus lotuta egon daitezke, batak helburu bat aurkeztu eta besteak helburu hori betetzeko plan bat ematen badu).

   Aurreko adibidean esan dugun bezalaxe, oraingoan ere, planifikazio mailan hainbat gauza dakizkigu. Horiek guztiak konputagailu batean sartzea ez da batere sinplea! Ulermena gai konkretu bati buruz mugatzen badugu, agian zerbait lor liteke konputagailuarekin, baina, oro har, edozeri buruz ulertzea eskatu nahi badiogu... jai dugu gaur egun!

Pragmatika testuinguruari dagokion informazioaz arduratzen da; hau da, berez linguistikoa ez den, eta igorpen linguistikoen prozesamenduan eta interpretazioan eragina duten informazioez arduratzen da. Bi atal bereiz daitezke hor:

- Diskurtsoaren ezagutza. Hemen lehenago igorri diren esaldien interpretazioez arduratzen dira. Anaforari dagozkion arazoak, eta denborari dagozkion ezaugarriak tratatzen dira beste batzuen artean. Atal honi buruzko informazio zabalagoa nahi duenak jo beza ondorengo liburuko 18. atalera (669-718. or.) Jurafsky D & Martin J.H *Speech and language processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, New Jersey 2000.

- Munduaren ezagutza. Hizkuntza bateko hiztunek elkarren artean komunikatzerakoan, munduari buruz duten ezagutza kontzeptual guztia hartu behar da kontuan. Horrelako ezagutzak esaldietan esplizituki adierazten ez den eta bistan den informazioa ulertzeko balio du.

  Faktore linguistiko eta ez-linguistikoen arteko erlazioak honako ondorio hauek ditu:

  - Ezagutza linguistikoa eta munduari dagokion ezagutza ezin dira independenteki tratatu, bataren eta bestearen artean jarraitutasun bat dago eta.
  - Bi ezagutza mota horiek errepresentatzeko sistemak bateragarria izan behar du. Hau da, komeni da bi ezagutza horietarako errepresentazio-sistema bakar bat egotea, hartara, haien informazioa aldi berean erabili ahal izango da. Alde batetik, kasu batzuetan munduari dagokion ezagutza oinarri linguistikoarekin eginiko sailkapenetan jasotzen da. Eta horrelako kasuetan, munduari dagokiona modu natural batez tratatzen da ezagutza linguistikoaren espezifikazio gisa. Bestalde, prozesamendu-sistema bat ezin daiteke osoa izan arrazoiketak egitera iristen den arte. Askotan, nahi den interpretazioak inferentzia edo dedukzio mailaren bat eskatzen baitu.

Hirugarren atala

HIZKUNTZA-TEKNOLOGIAKO PRODUKTUAK

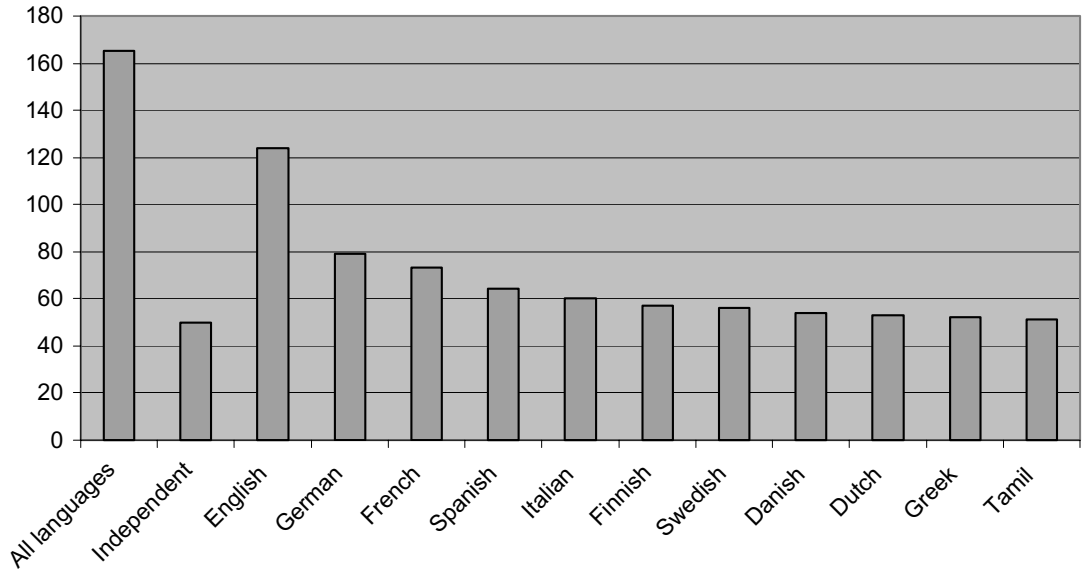
Egunetik egunera, modu elektronikoan gure esku dagoen testu-masa handitzen doa, gainezka ere egiteraino batzuetan, Interneten edo posta elektronikoan bilatu nahi dugun testu zati hori nola aurkitu ez baitugu sarritan asmatzen. Horren ondorioz, ezinbesteko bihurtzen zaigu hizkuntza informatikoki lantzeko aplikazioak erabiltzea. Zalantzarik gabe, hizkuntza-teknologiak funtsezkoak dira *Informazio eta Komunikazioaren Gizartea* esaten dugun horretan.

LNPren ia 50 urteko historian gorabehera handiak izan dira. Helburu liluragarriak lortzear zeudela uste zen une euforikoei, belarriak jaitsi eta helburu apal baina eskuragarriagoetara mugatzeko uneak jarraitu zaizkie behin baino gehiagotan. Hala nola, erabateko itzulpen automatikoa konputagailuen eskutik etorriko zela aurreikusi zuten 1954an Georgetown-eko Unibertsitatean. Alabaina, 1966an itzulpen automatikorako diru-iturri ofizial guztiak itxi egin ziren Amerikako Estatu Batuetan, ALPAC txosten ezagunak horrela gomendatu eta gero. Aurrerago, 1980 inguruan, adimen artifizialeko teknika berrien eskutik, konputagailuak hizkuntza arruntaz —lengoaia naturalean— programatu ahal izango genituela agindu zitzaigun. Gaur egun ahaztuta daude horrelako ametsak.

Gaur egun, batetik, hizkuntzaren egitura eta erabileraren zailtasuna aitortzen dugu, ez direla hasieran uste bezain sinpleak; eta bestetik, helburu utopiko haiek baztertuta helburu apalagoa duten baina komertzialki bideragarriak diren produktu asko merkaturatu dira. Tresna mugatuak dira, eta beti errore maila batekin, baina, hala ere, laguntza ederra ematen digute. Alde batetik, ekonomikoki errentagarriak direlako (merkeagoa da erroreak dituen itzulpen-zirriborro bat zuzentzea, testu osoa bere osotasunean itzultzea baino); eta bestetik, tresna horiei esker gizakien arteko komunikazioa hobetu egiten delako (adibidez, testu zuzenago eta zehatzagoak sortuz, edo telefono bidez beste hizkuntza darabilen pertsona batekin hitz egitean hitzak itzuliz)

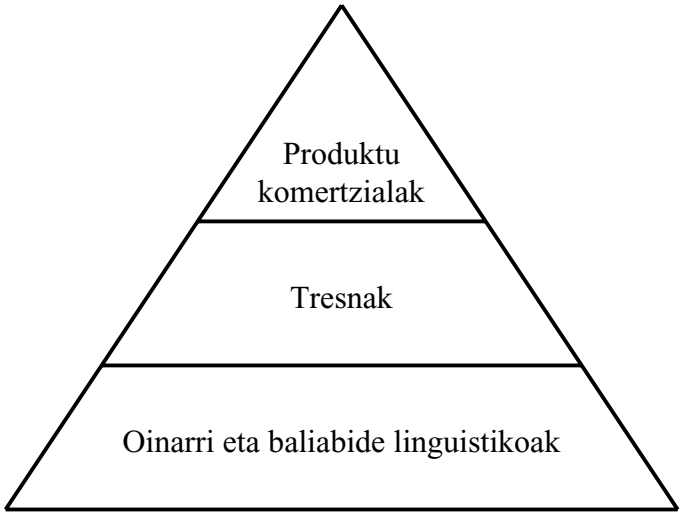
Sarreran aipatu dugun bezala, gaur egun badira zenbait hizkuntza-aplikazio eskuragarri: ortografia-zuzentzaileak eta estilo-zuzentzaileak, hiztegi-kontsultak on-line, itzulpen-laguntzak, Interneterako bilatzailea, Hizketa testu bihurtzen duten sistemak, testuak irakurtzen dituztenak, bigarren hizkuntza ikasteko sistemak eta abar.

Baina horrelako sistema gehienak ingeleserako sortu dira, ez gainerako hizkuntzetarako. Gainerako hizkuntzek ahalegin handia egin behar dute atzean ez gelditzeko. Are gehiago euskarak eta euskara bezalako hizkuntza txikiek. *Natural Language Software Registry* zerbitzuak Interneten duen orria aztertzen badugu (http://registry.dfki.de) egun hizkuntzak lantzeko erabilgarri diren 167 programaren berri jasoko dugu (ikus 11. irudia). Horietatik % 75 ingeleserako erabilgarri dira, eta % 30 bakarrik erabil daitezke edozein hizkuntzatarako. Merkatuan aurki daitezkeen aplikazio gehienek hizkuntza “handiak” dituzte helburu, ingelesa, batik bat, baina baita, bigarren maila batean bada ere, frantsesa, alemana eta espainiera, besteak beste.



11. irudia. NLSR katalogoko produktuen aplikagarritasuna hainbat hizkuntzatan

Konputagailuek hizkuntza gizakiok ulertzen dugun moduan ulertuko duten eguna urrun da oraindik, baina horrek ez du esan nahi aplikazio interesgarri eta oso baliagarriak egin ezin direnik.



12. irudia. Produktuak eta produktuak sortzeko ordena

Aplikazio horien garapenerako, ordea, oinarri sendo batetik abiatu beharra dago. Oro har, hizkuntza-teknologien egitura piramide moduko batez irudika dezakegu (ikus 12. irudia). Piramide horren oinarrian ingeniaritza linguistikoan lan egiteko beharko ditugun oinarrizko baliabideak daude. Baliabide horiei esker, tresnak garatzeko moduan izango gara, eta behin tresnak garatuta, ingeniaritza linguistikoaren hainbat arlotan lan egiteko moduko produktu komertzialak kaleratu ahal izango ditugu.

Hori kontuan hartuta, LNPren barruan azaltzen diren produktuetan hiru multzo nagusi egingo ditugu: lehenengoan, hizkuntzalaritzaz edo informatikaz gutxi dakien erabiltzaile arruntarentzat salgai diren **aplikazioak** sartuko ditugu, hizkuntza automatikoki tratatzeak zer helburu praktiko dituen azalduz; bigarrenean, aplikazio horiek sortuko badira zer-nolako azpiegitura behar den azaltzen saiatuko gara, hizkuntza-softwarea sortzen dutenentzako **tresnak**, produktu berriak garatzeko baliagarriak direnak; eta bukaeran aztertuko ditugu edozein aplikazio edo tresna garatzeko behar-beharrezkoak diren **hizkuntza-baliabide** eta **oinarri linguistikoak.**

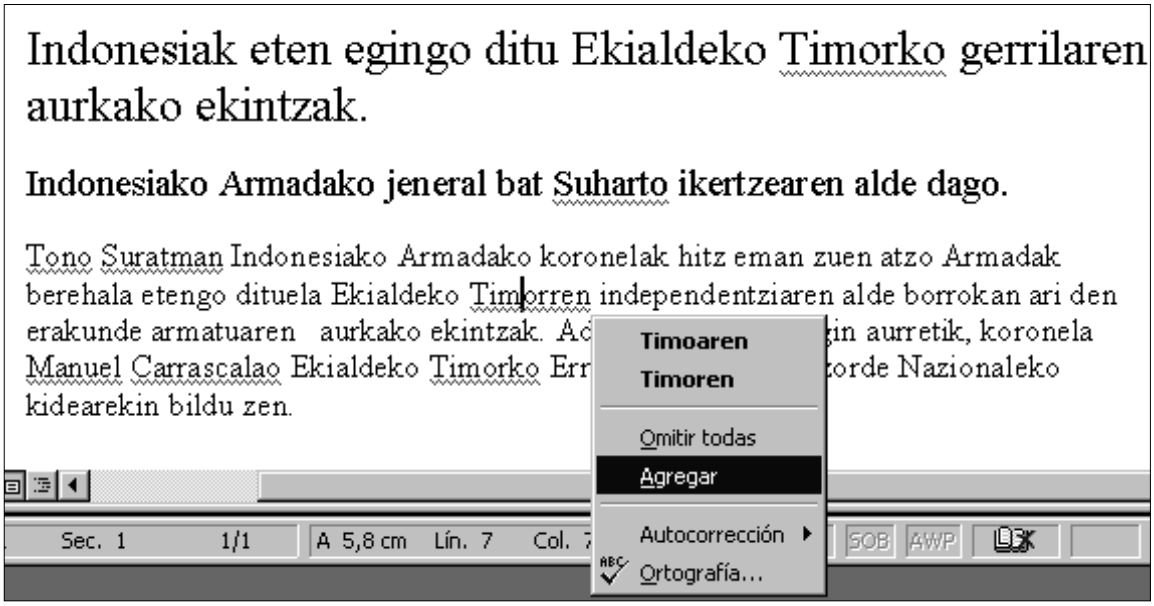
6 Aplikazioak erabiltzaile arruntarentzat

6.1 Testuak editatzeko eta ulertzeko laguntzak

Egun badira testu-egileari eskaintzen zaizkion laguntza bereziak. Ikus ditzagun orain zein diren garrantzitsuenak.

6.1.1 Ortografia-zuzentzailea

Bere helburua testuen akats ortografikoak detektatu eta zuzentzea da. Horretarako kontrolatzen du ea hitz



13. irudia. Ortografia-zuzentzailea

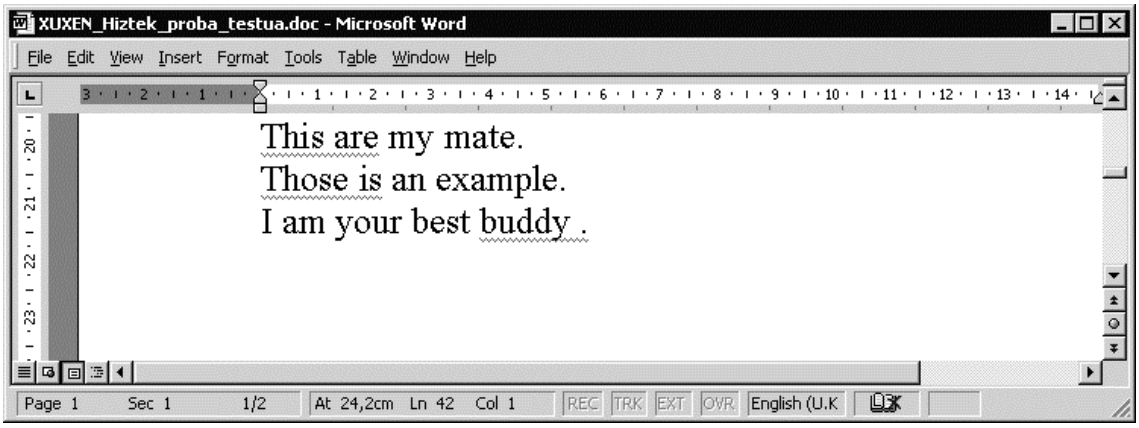
bakoitza posible ote den, eta horrelakorik ez bada antzekoenak diren hitz posibleak proposatuko ditu, baina ez du errore sintaktiko edo semantikorik detektatzen. Gaur egun hizkuntza askotarako aurki daitezke, eta 1994tik euskararako ere bai: Xuxen, euskararako zuzentzaile ortografikoa.

Xuxen-ek euskara batuaren erabilera bultzatzen duenez (Euskaltzaindiaren azken erabakiak barne), ez ditu ontzat ematen forma dialektalak, nahiz eta batzuk ezagutu eta beren ordezko estandarra proposatu. Adibidez, **haundi* zuzentzeko *handi* proposatuko digu ; edota**iharduten* hitza zuzentzeko *jarduten*.

Gure inguruko beste hizkuntzetako zuzentzaileekin alde nabarmen bat dauka, euskararen atzizki ugarien erabilera dela eta, ez baita oso bideragarria euskaraz hitz posible guztien zerrenda sortzea. Hitzak morfologikoki analizatu egin behar dira ontzat eman ahal izateko. Hitz berri bat hiztegian sartu ahal izateko euskarako zuzentzaileak bere kategoria morfologikoa galdetzen du geroago berarekin sortutako hitz flexionatu guztiak ere onartu ahal izateko.

6.1.2 Gramatika- eta estilo-zuzentzaileak

Hauek testuingurua kontuan hartzen dute. Adibidez, “nik joan naiz” esaldia prozesatuz gero, ortografia-zuzentzaileak ez luke errorerik salatuko, hiru hitzok isolatuta posible baitira, baina sintaxi-zuzentzaileak testuinguru horretan “nik” hitza gaizki dagoela salatuko luke eta “ni” izan beharko lukeela proposatu.



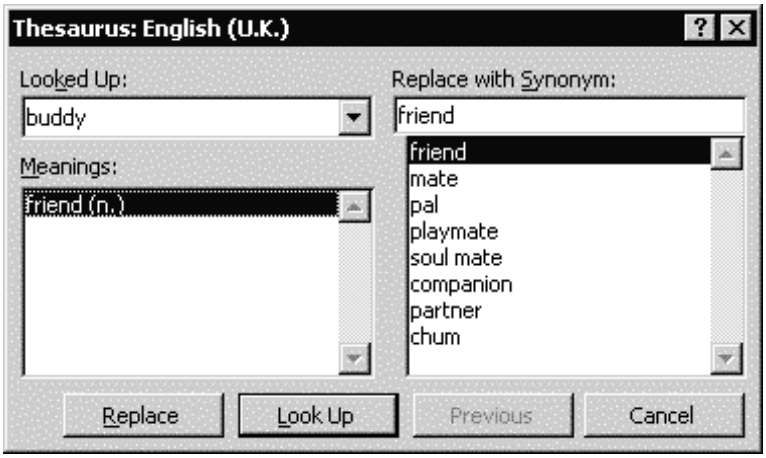
14. irudia. Gramatika- eta estilo-zuzentzailea

Estilo-akatsak ere salatzen dituzte; adibidez, perpausaren azken hitza eta bukaerako puntua zuriune karaktere batez banaturik agertzen direnean. Errore hori azaltzen da 14. irudiko azken esaldian.

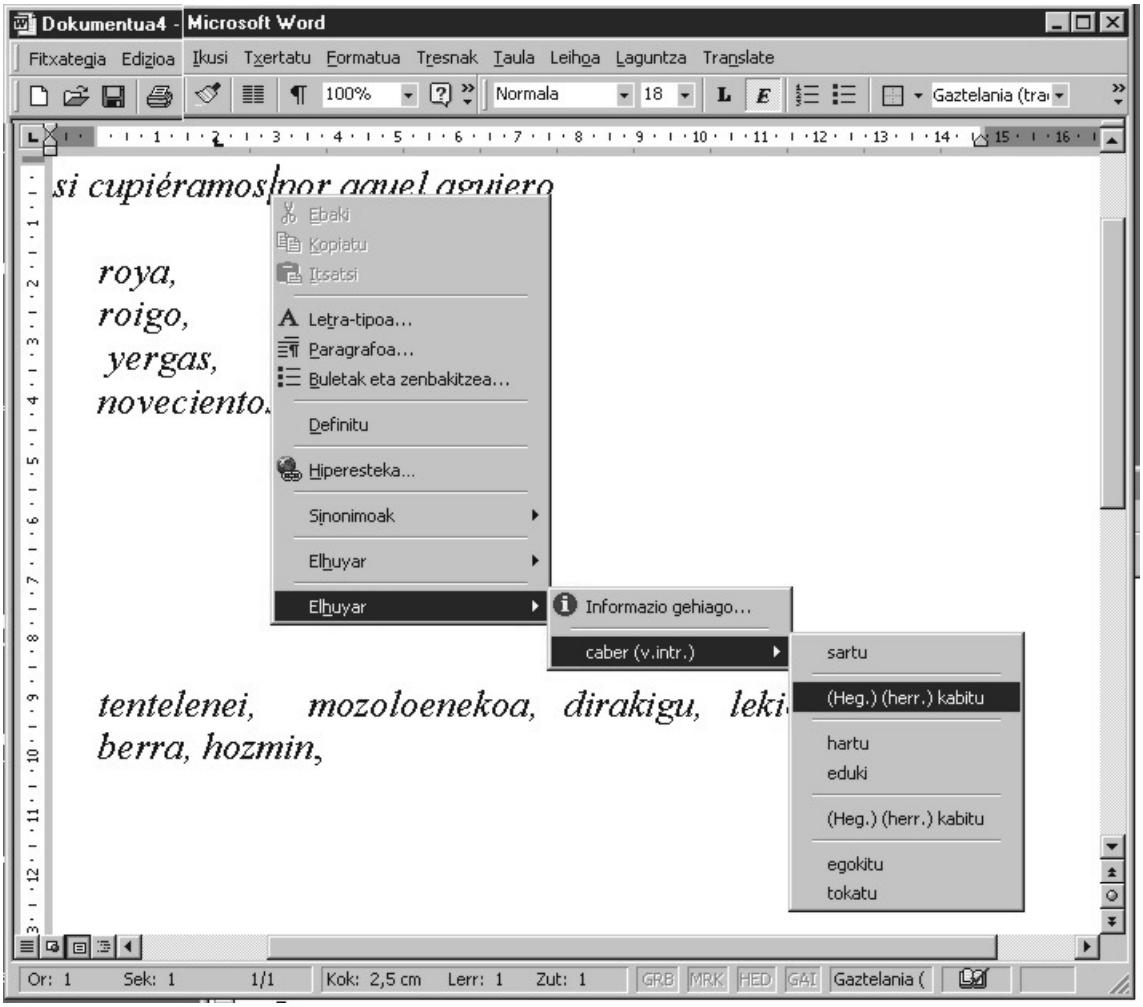
6.1.3 Hiztegi-kontsultarako laguntzak

Era honetako laguntza lexikaletan edozein hitz kontsulta daiteke hiztegi batean, kasu batzuetan testu-prozesaketarako programatik atera gabe hitzaren gainean klik eginda. Horrela posible da lortzea sinonimoak, beste hizkuntza bateko baliokideak, taxonomikoki konkretuagoak edo orokorragoak diren antzeko hitzak ere, thesaurusa kontsultatuz (adibidez: *intsektu* hitzetik orokorragoa den *animalia* edo konkretuagoak diren *inurri*, *euli*..).

On-line moduko hiztegien ingelesezko adibide gisa, 15. irudian ikus dezakegu nola lortu *buddy* hitzaren esanahia bere sinonimoen bidez.



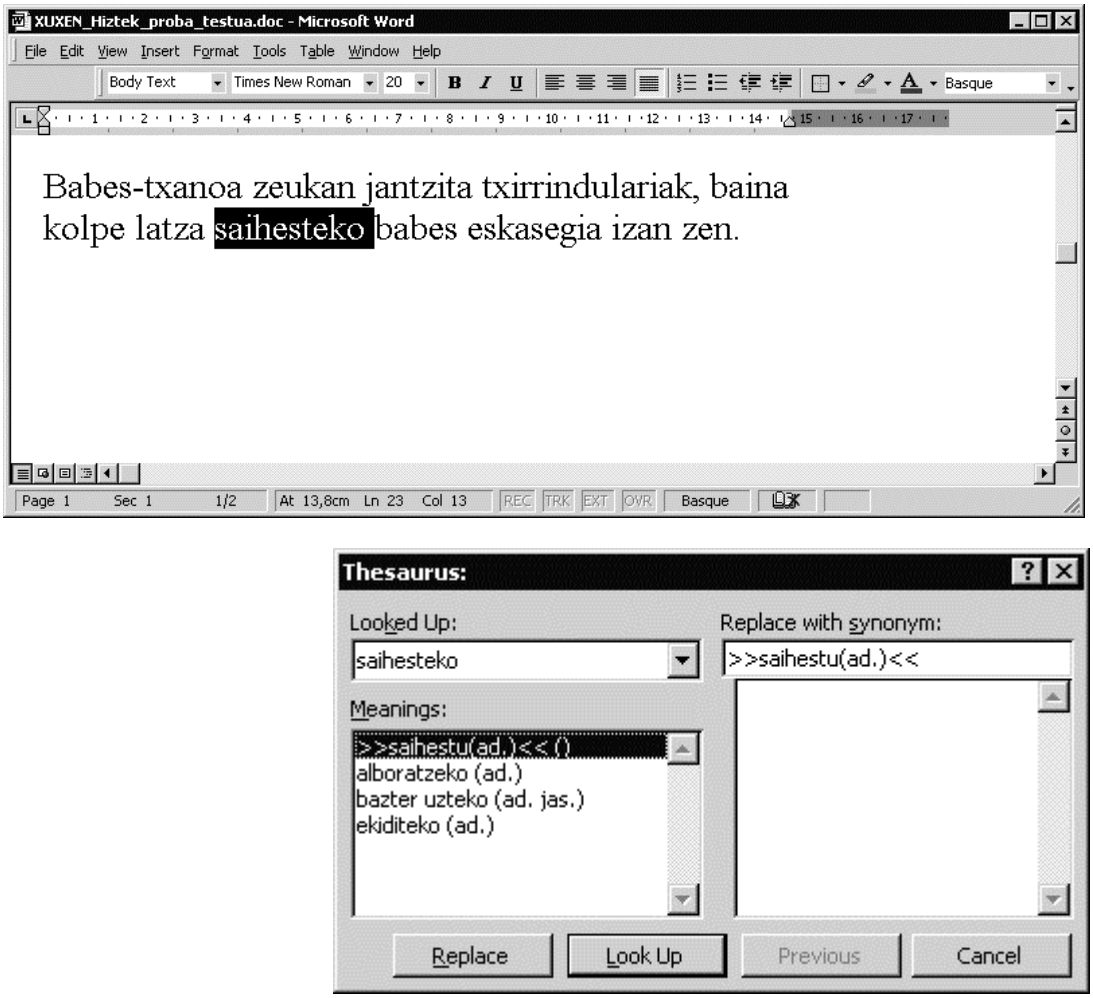
15. irudia. Nola lortu *buddy* hitzaren esanahia ingelesezko sinonimoen bidez



16. irudia. Hiztegi elebiduna on-line eta lematizazioan oinarritua

Hiztegi elebidunen adibide gisa 16. irudian ikus daiteke nola kontsultatu erdarazko *cupiéramos* hitza Gaztelania-Euskara Elhuyar Hiztegian (*sartu, kabitu, hartu, eduki, egokitu, tokatu*). Kontuan hartu galdera hori ez dela erraza paperezko hiztegi batean egiteko, *cupiéramos* hitza kontsultatzeko oso urruti dagoen *caber* hitza bilatu behar dela ez badakigu behintzat. Hitzaren lema, erroa, lortu behar da hiztegi-kontsulta bera egin aurretik.

Hitzen lematizazioa egiten da sistema horretan kontsulta egiterakoan, baina sorkuntza morfologikoa ere egiten da UZEI Sinonimoen Hiztegian kontsultatzean. Horrela, *saihesteko* hitzaren sinonimoak eskatzen ditugunean, atzizki berekin azalduko zaizkigu baliokideak (*alboratzeko, bazter uzteko* eta *ekiditeko*) (Ikus 17. irudia).

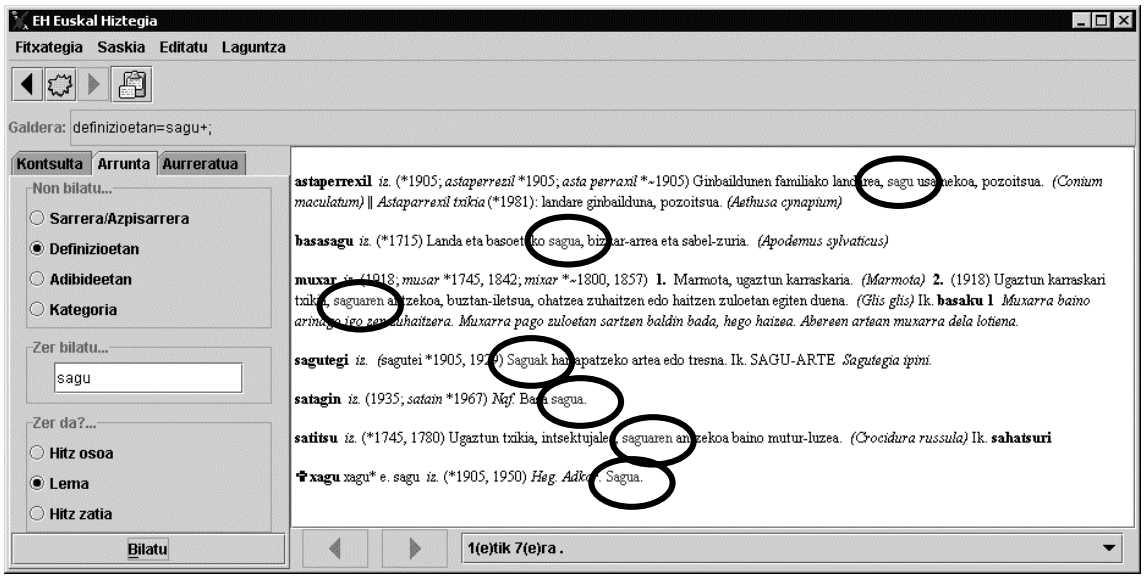


17. irudia. Sinonimo-hiztegia on-line eta lematizazioan oinarritua

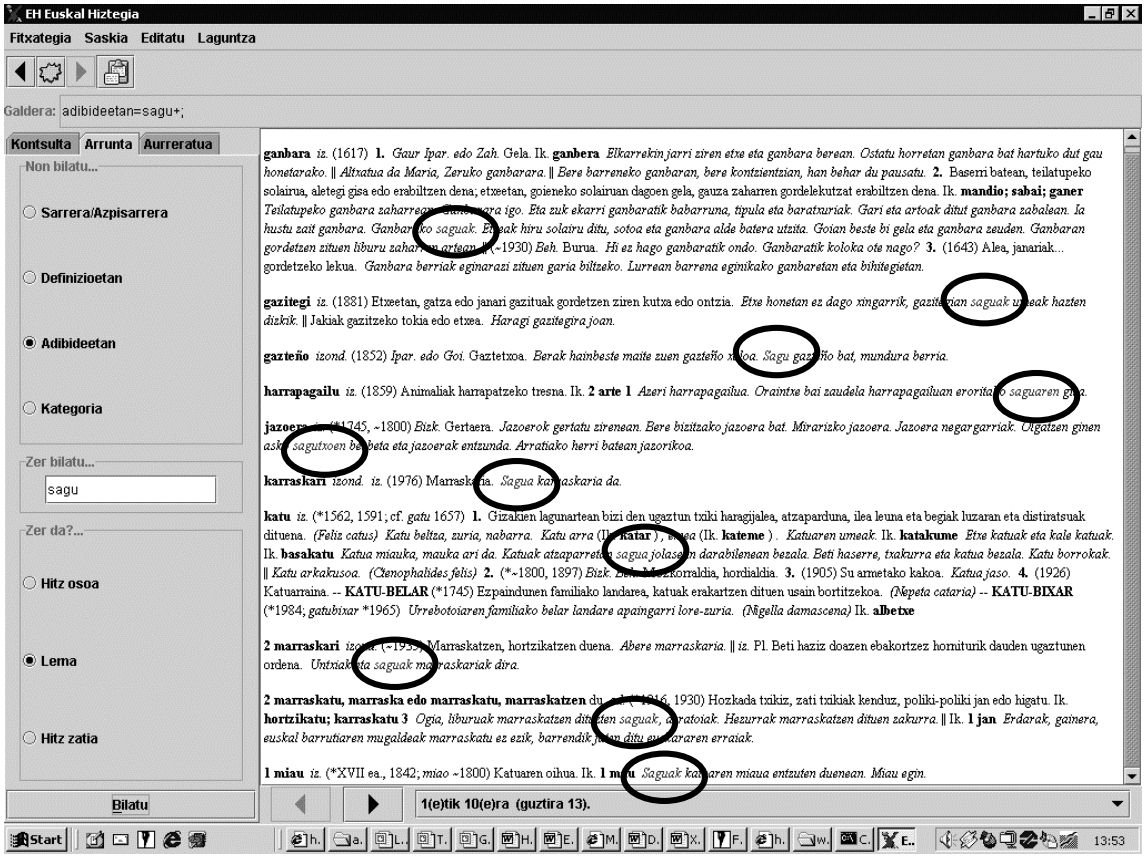
Erabilera finagoak egin daitezke, hala ere, hiztegi baten kontsultan eta kalitatezko dokumentazioa sortu nahian. Harantzago doa *Euskal Hiztegia*ren bertsio elektronikoa. Kasu honetan, hitz baten definizioa eta erabilera-adibideak lortzeaz gain (orain arteko beste edozein hiztegitan bezala), aukera dago ideia batetik abiatuta hitzaren bila joateko. Adibidez, posible da bilatzea *sagu* hitza hiztegiko zein sarreratan erabili den definizioan (*astaperrexil, basasagu, muxar, sagutegi, satagin, satitsu, xagu*) (ikus 18. irudia) edo adibideetan (*ganbara, katu, marraskari, marraskatu, harrapagailu*...) (ikus 19. irudia).

Lematizazioa erabiltzen denez, erantzunak ez dira nahasten *saguzar* hitzari dagozkion emaitzekin. Hori gertatuko litzateke bilaketan *sagu** jarri izan bagenu, hau da, sagu hitz-hasiera bilatuz. Lematizazioari esker, emaitzak zehatzagoak dira, zabor gutxiago azaltzen zaigu. *Saguzar* hitzarentzat agertuko liratekeen emaitzak 20. irudian erakusten dira, baina horiek ez dira azaltzen *sagu* lema erabiltzen dugunean.

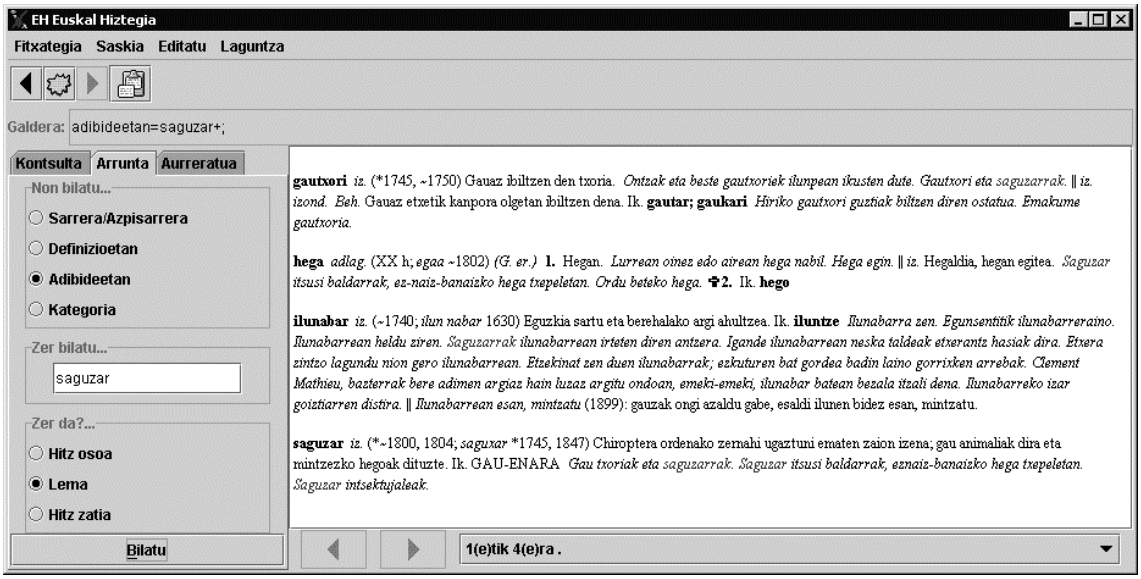
Beste aukera batzuk ere badira: edozein hitzetan klik eginez gero bere definizioa erakusten da. Kategoria sintaktikoaren arabera ere bila daiteke, atzizkiaren arabera...



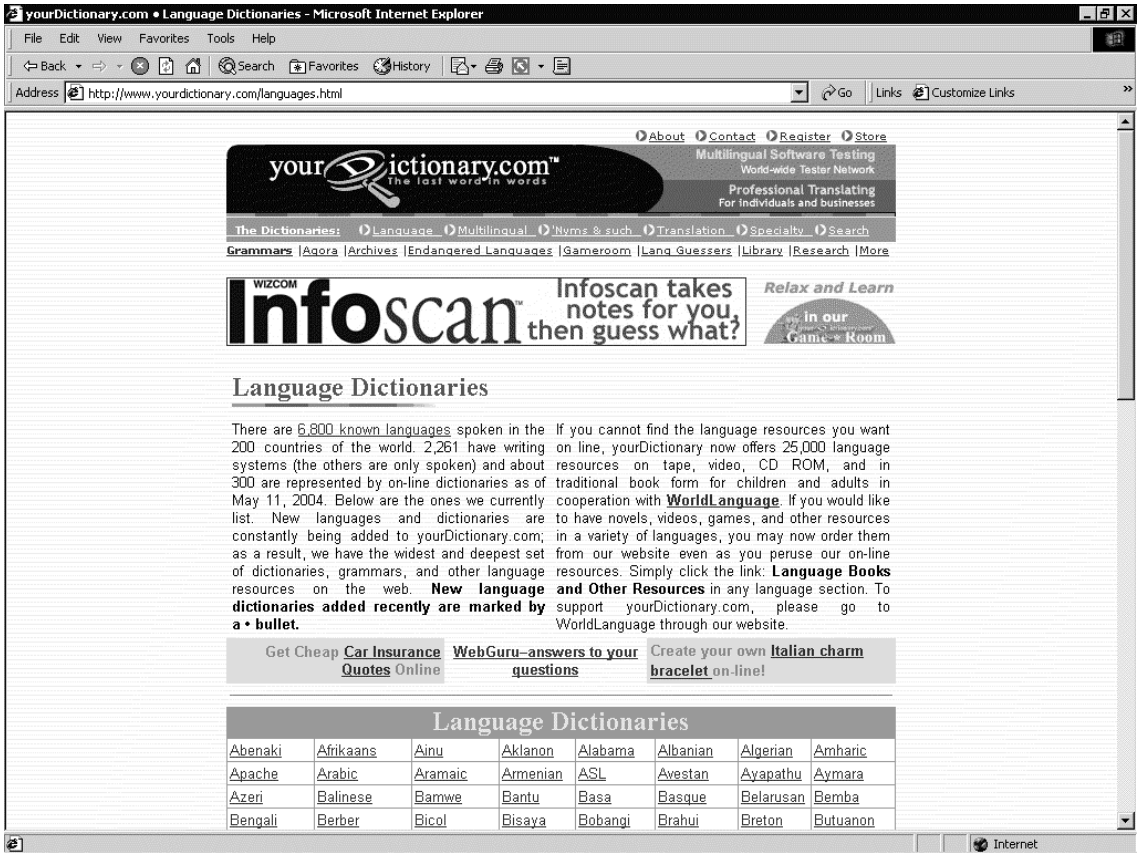
18. irudia. Euskal Hiztegiaren bertsio elektronikoa. *Sagu* lema zein definiziotan?



19. irudia. Euskal Hiztegiaren bertsio elektronikoa. *Sagu* lema zein adibidetan?



20. irudia. Euskal Hiztegiaren bertsio elektronikoa. *Saguzar* lema zein definiziotan?

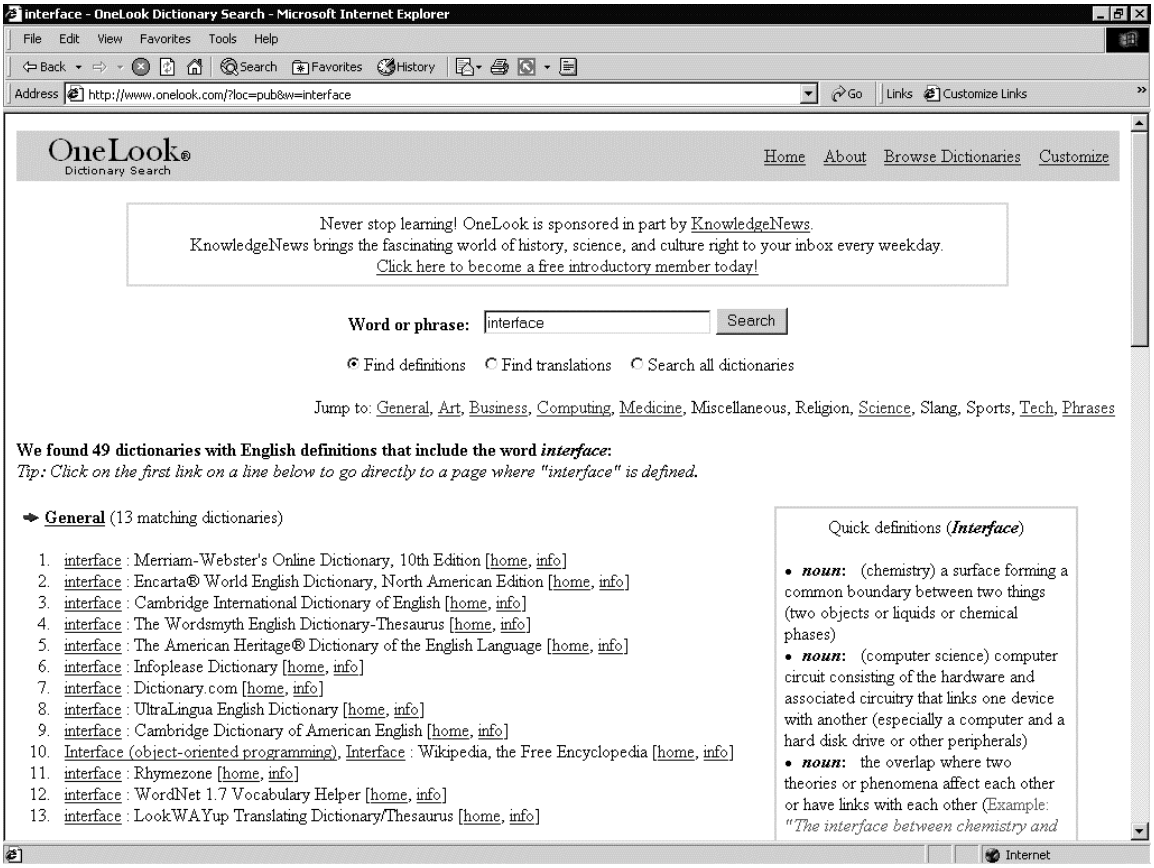


21. irudia. 300 hizkuntzatarako hiztegiren berri yourdictionary.com web gunean

Hiztegi-kontsultarako laguntzen atal honi bukaera emateko datu batzuk azalduko ditugu. www.yourdictionary.com helbidera jotzen badugu, munduan hiztegi elektronikoen garapenak hartu duen neurria ikusiko dugu:

- 300 hizkuntzatarako hiztegi elektronikoen berri bildu da web gune horretan (ikus 21. irudia, kontuan hartu munduko hizkuntzak 6.800 direla, eta horietarik 2.261ek bakarrik dutela adierazpide idatzia).
- Euskararako 9 hiztegi aipatzen dira bertan. Erreferentzia gehiago aurkituko dugu www.hiztegia.net gunean baina. Honetan gutxienez 18 hiztegi orokor eta 29 berezitu azaltzen baitira.
- Gaztelaniarako 100 baino gehiago.
- Ingeleserako 870 baino gehiago. Web gunean bertan aukera eskaintzen da hitz bat hiztegi horietan guztietan batera galdetzeko (ikus 22. irudia). Hiztegi batzuk berezituak dira eta, adibidez, badira 6 hiztegi hizkuntzalaritzari buruz:
  - Dictionary of English Usage
  - Dictionary of Phonasthemes
  - A Glossary of Translation and Interpreting Terminology
  - IPA symbols with downloadable recordings
  - Lexicon of Linguistics (the latest English theoretical terminology)
  - Linguistic Glossary (Summer Institute of Linguistics)

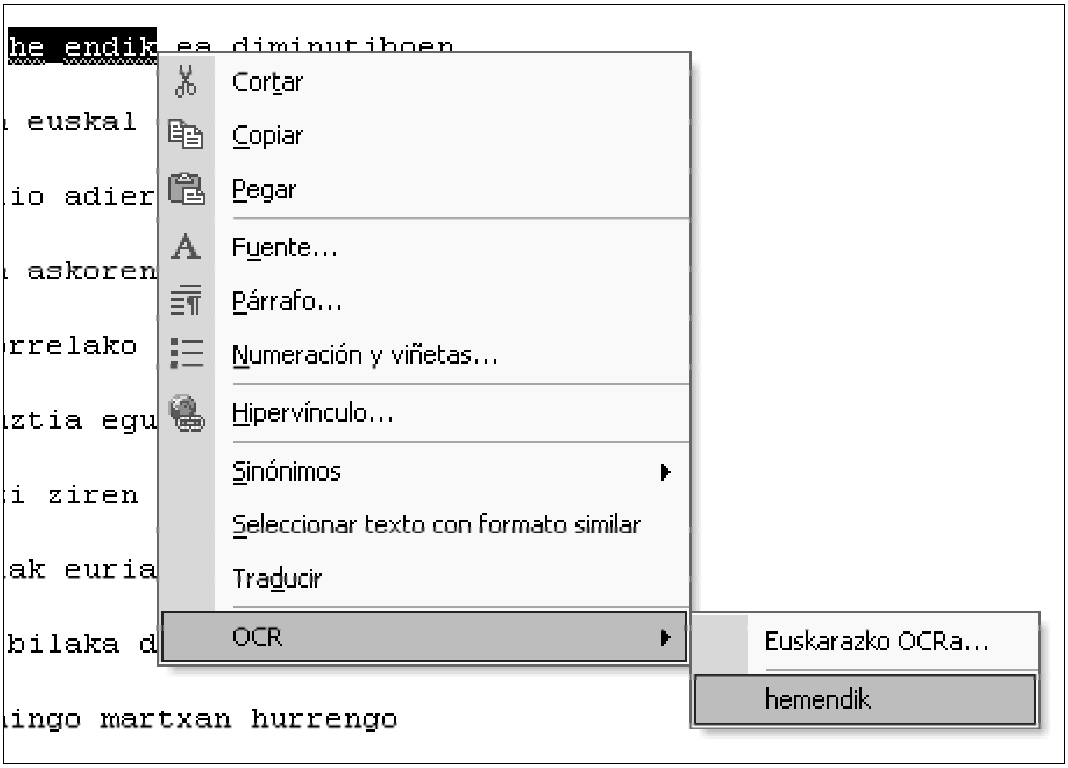
  eta 15 konputagailuei buruz:
  - Hutchinson Dictionary of Computers, Multimedia and the Internet
  - Denis Howe's FOLDOC
  - CCI High Tech Dictionary
  - Comprehensive Technology Glossary
  - Computer User High Tech Dictionary
  - Computerese Jargon Monitor
  - Dictionary of Computing Terms (HTML)
  - A Dictionary of Storage Networking Terminology
  - Elsevier's Dictionary of Computer Science and Mathematics
  - Elsevier's Dictionary of Personal and Office Computing
  - Fujitsu Electronics Dictionary
  - A Glossary of Computer Oriented Abbreviations and Acronyms
  - Glossary of Memory Terms
  - Digital Compression Glossary Of Terms
  - Computer Buyers' Glossary



22. irudia. Ingeleserako 870 hiztegi eskura kontsulta bakar batean

6.1.4 Testuak egoki digitalizatzeko OCR programak

Scaner baten bitartez testu bat digitalizatzen dugunean errore tipiko batzuk gertatu ohi dira, adibidez "h" karakterea "n" bihurtzea. Badira laguntza-programa bereziak ezagutza linguistikoa erabiliz errore horiek leuntzen dituztenak. Euskararen kasuan ere bada horrelako laguntzarik. Elekak kaleratutako OCR1.1Euskaraz programa zuzentzaile ortografiko berezia da. OCR programek egin ohi dituzten errore tipikoak detektatu eta zuzentzen ditu, grafikoki itxura beretsukoak diren erroreak, hain zuzen. Esate baterako "lo" batzuetan "b" letrarekin nahas daiteke. Errore hori ez da errore arrunta testuak tekleatzean. Omnipage programaren barruan edo MSword programa barruan integratuta erabiltzen da.



23. irudia. OCR bidez lortutako testuetarako zuzentzaile berezitua

6.1.5 Testu eleaniztunak editatzeko laguntzak

Testu eleaniztunak lantzeko adibide gisa Trados eta Wordfast aipatu behar dira. Prozesadore zabalduenetan integratzen diren programa hauek glosategi, hiztegi eta itzulpenen berrerabilerarako laguntzak eskaintzen dituzte.

6.2 Testu-masa handiak tratatzeko edo kudeatzeko aplikazioak

Modu elektronikoan gure esku dagoen testu-masa handia da. Batzuetan testu-masa erraldoi hori modu egituratuan dago, testu guztiak datu-base dokumental gisa antolatu direlako, baina gehienetan arazoa sortzen zaigu dokumentu andana eta egituratu gabeak tratatu nahi ditugunean. Eta horrela, esate baterako, batzuetan ez dugu asmatzen Interneten aurkitu nahi dugun informazioa, edo aspaldian posta elektroniko bidez bidali ziguten mezu berezi hura.

Testu-masa handiak tratatzeko edo kudeatzeko arlo nahiko berriari *jakintzaren kudeaketa* (*Knowledge-management*) hasi zaio esaten. Horko aplikazio nagusiak bi dira:

- Dokumentuen berreskurapena (IR, *Information Retrieval*) eta
- Informazio-erauzketa edo datuen erauzketa dokumentuetatik abiatuta (IE, *Information Extraction*)

Hala ere, badira beste aplikazio mota berezituago batzuk ere:

- Laburpen automatikoa (S*ummarization*)
- Dokumentu-sailkatzaileak
- Dokumentuak bideratzea (*Routing*)
- Dokumentuak multzokatzea (*Clustering*)
- Dokumentuen iragaztea (*Filtering*)
- Testu-sorkuntza automatikoa

6.2.1 Dokumentuen berreskurapena (IR, *Information Retrieval*)

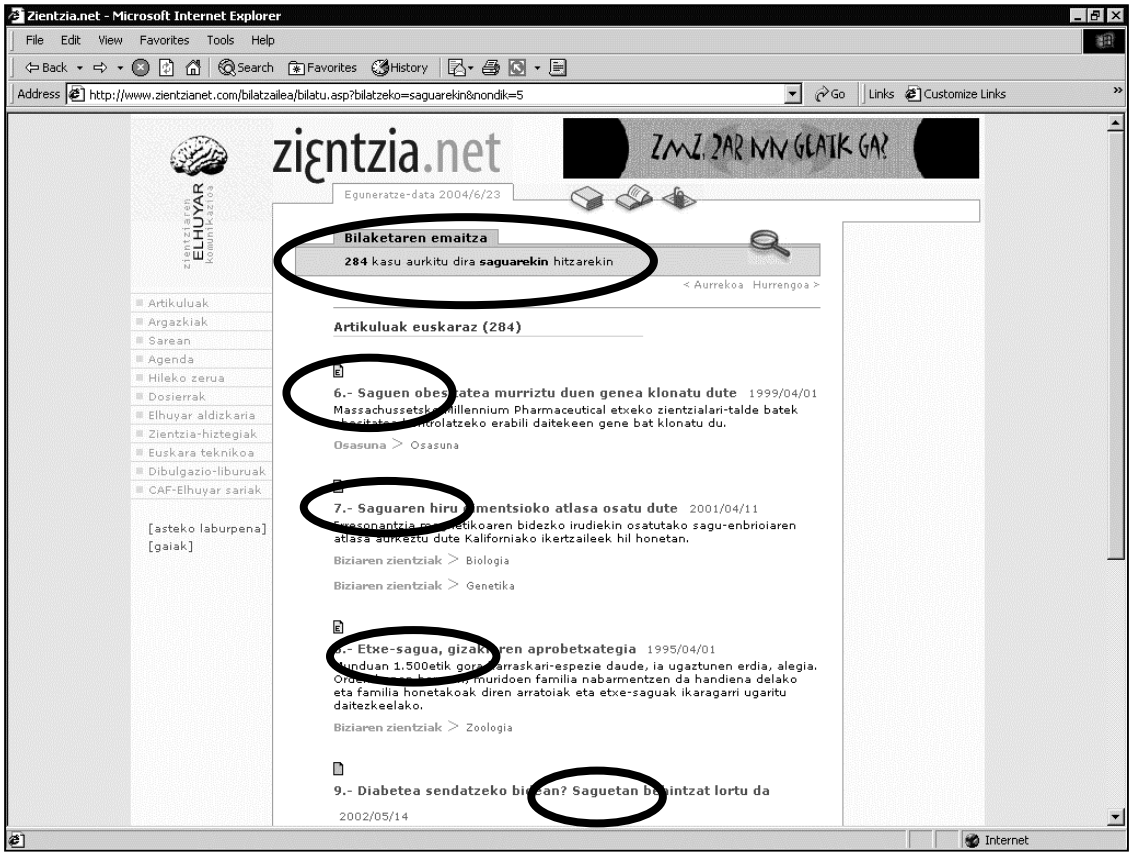
Aplikazio honen helburua da hainbat eta hainbat dokumenturen artean bakar bat (edo batzuk) hautatzea, kontzeptu bat edo informazio bat daukana. Noski, adibide tipikoena Interneterako bilatzaileena da, esaterako, Google (www.google.com).

Programa hauek barruan bi modulu dauzkate beti: bata, *modulu indexatzailea,* eskura dituen dokumentuak aztertzen dituena barruko hitz edo kontzeptuekin indizeak sortzeko; eta bestea, *modulu bilatzailea,* bilaketak azkarrago egitea ahalbidetzen duena. Interneteko bilatzaileen kasuan, modulu indexatzaileak etengabe daude martxan, web gune berriak detektatzen, analizatzen eta indizeak eguneratzen.

Nazioartean hainbat bilatzaile dira aipagai, eta euskararen tratamenduan ere egin izan dira zenbait ekarpen. Halabeharrez egin behar izan dira, euskarazko testuetan hitz osoak bilatzea ez baita oso praktikoa, sarritan hitzetan atzizkiak azaltzen baitira; eta hitz-hasierak bakarrik bilatzen baditugu, horrelaxe hasten diren beste hitz luzeagoei dagozkien emaitzak ere azalduko zaizkigu, emaitzen kalitatea zapuztuz. Adibidez, *ero* hitza duten dokumentuak bilatu nahi baditugu, *eroari*, *eroekin*, *eroengana* hitzak dituzten dokumentuak ere detektatu nahi ditugu; konponketa bat litzateke “ero” letrekin hasten diren hitz guztiak detektatzea (*ero** bilatzea), baina horrelakoetan *erosotasun*, *erosi*, *erosten*, *eroale*... hitzen aipamenak dituzten dokumentuak ere jasoko ditugu, eta horrelakorik ez dugu nahi, azken horien

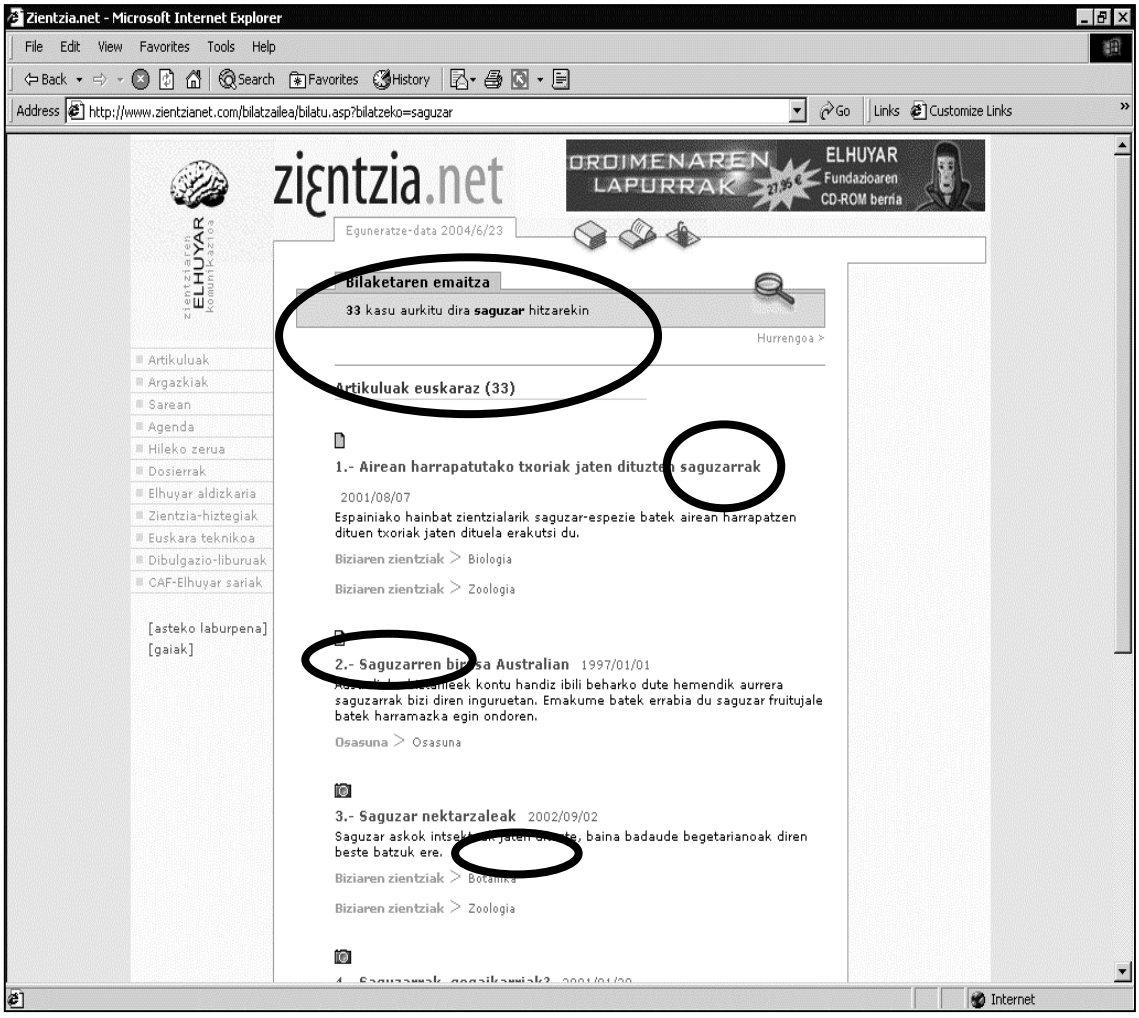
erreferentziak agertzen badira, benetan bilatzen ditugunekin nahastatuta agertuko direlako. Beraz, ahal dela, lematizazioan oinarritutako bilaketak egin beharko ditugu euskarazko dokumentuak atzitzeko.

- Ametzagaina taldeak kaleratutako *Kapsula* softwarea, euskarazko dokumentu-baseen kudeaketara zuzenduta dago. Diana Teknologia enpresak, enpresa-kudeaketan erabili behar izaten diren testu, inprimaki eta, oro har, ezagutza guztia ustiatzeko tresnak lantzen ditu (Xerka), hizkuntza-teknologiako tresnak integratuz. Enpresa-kudeaketa horretan gizakiaren eta sistemaren arteko elkarrekintza ahalbidetzen duten sistemak dira; erakundeetan biltzen den ezagutza erraldoia ustiatzeko tresnak dira, baita ikasketarako edota erabaki-hartzeak errazteko ere.



24. irudia. ZientziaNet: lematizatzailea darabilen dokumentu-bilatzailea

- IXA taldearen EUSLEM lematizatzailea zenbait web gunetan integratu da. Adibidez, Berria egunkariaren hemerotekan eta ZientziaNet-en. 24. irudian ikus daiteke zer lortu den ZientziaNet-en "saguarekin" hitza galdetuta. Galderako hitza lematizatu ondoren programak *sagu* lemari dagozkion dokumentuak bilatzen ditu. 284 aurkitu ditu, baina horien artean ez dira azaltzen *saguzar* hitza (edo lema) dauzkaten 33 dokumentuak (ikus 25. irudia). Galdera "*sagu**" jarri izan bagenu, *saguzar*renak eta *sagu*renak, denak batera, azalduko ziren.



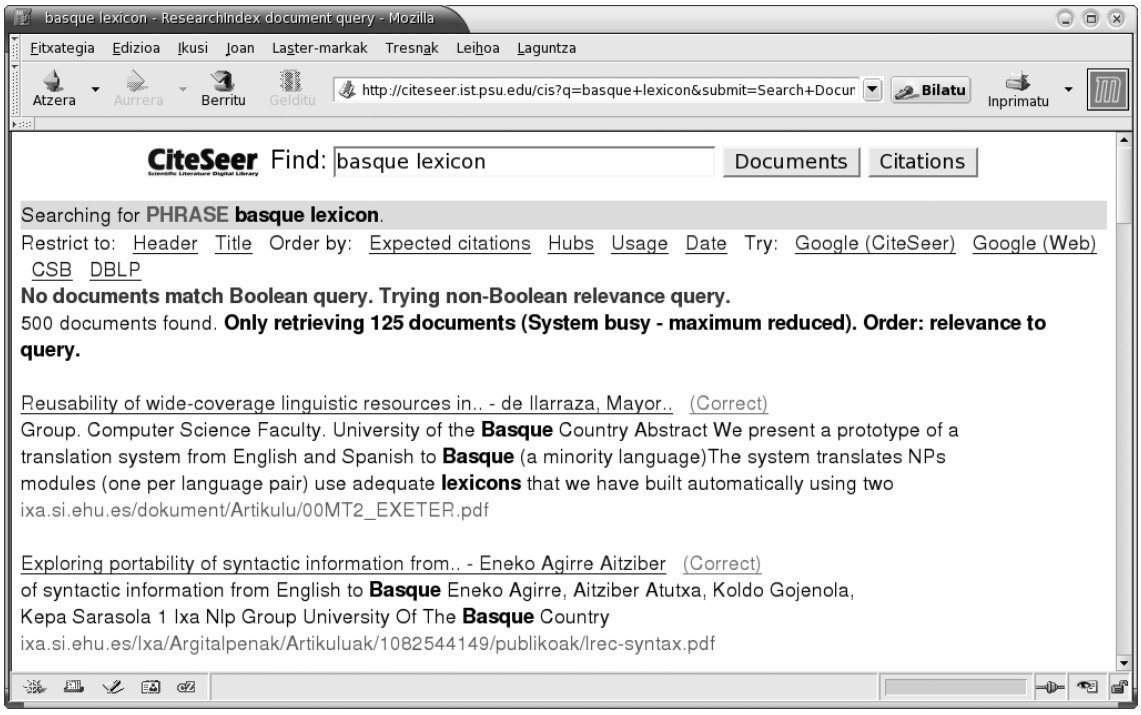
25. irudia. 33 artikulu *saguzar* lemarako, *sagu* bilatzean azaltzen ez direnak

6.2.2 Informazio-erauzketa (IE, *Information Extraction*)

Aplikazio mota honetan datuen erauzketa egiten da dokumentuetatik abiatuta. Funtsa dokumentu bakoitzean dagoen informazio esanguratsuaren detekzioa da: **Entitate, erlazio** edota **gertaer**ei buruzko informazioa ateratzea **domeinu mugatu** bateko dokumentuen artean. Gehienetan dokumentu bakoitzetik ateratako datuekin fitxa bat betetzen da eta datu-base batean integratzen da gero, horrela datu-base oso bat automatikoki betetzen delarik.

Gaur egun, hainbat domeinutan lortu dira aplikazio praktikoak: administrazioa, komunikabideak, medikuntza, salerosketak eta arlo militarra.

Adibide tipiko bat izan daiteke Citeseer (citeseer.nj.nec.com). Berez, Interneteko zerbitzu honek zientzia-artikuluak bilatzeko tresna eskaintzen du, baina atal honetan interesatzen zaiguna da jakitea sistemak era horretako artikuluak Interneten automatikoki bilatu, jaso eta sailkatzen dituela (egilea, izenburua, gaia...).



26. irudia. Citeseer: zientzia-artikuluak bildu eta eskaintzen dituen sistema

Informazio-erauzketa egiten duen sistema baten barruan modulu lagungarri hauek izaten dira: kontrol-karaktereen eta zati ez-interesgarrien iragazketa, morfologia, etiketatzailea, entitateen ezagutzailea eta sailkatzailea, azaleko sintaxia, patroien azterketa, inferentzia eta datu-basea osatzekoa.

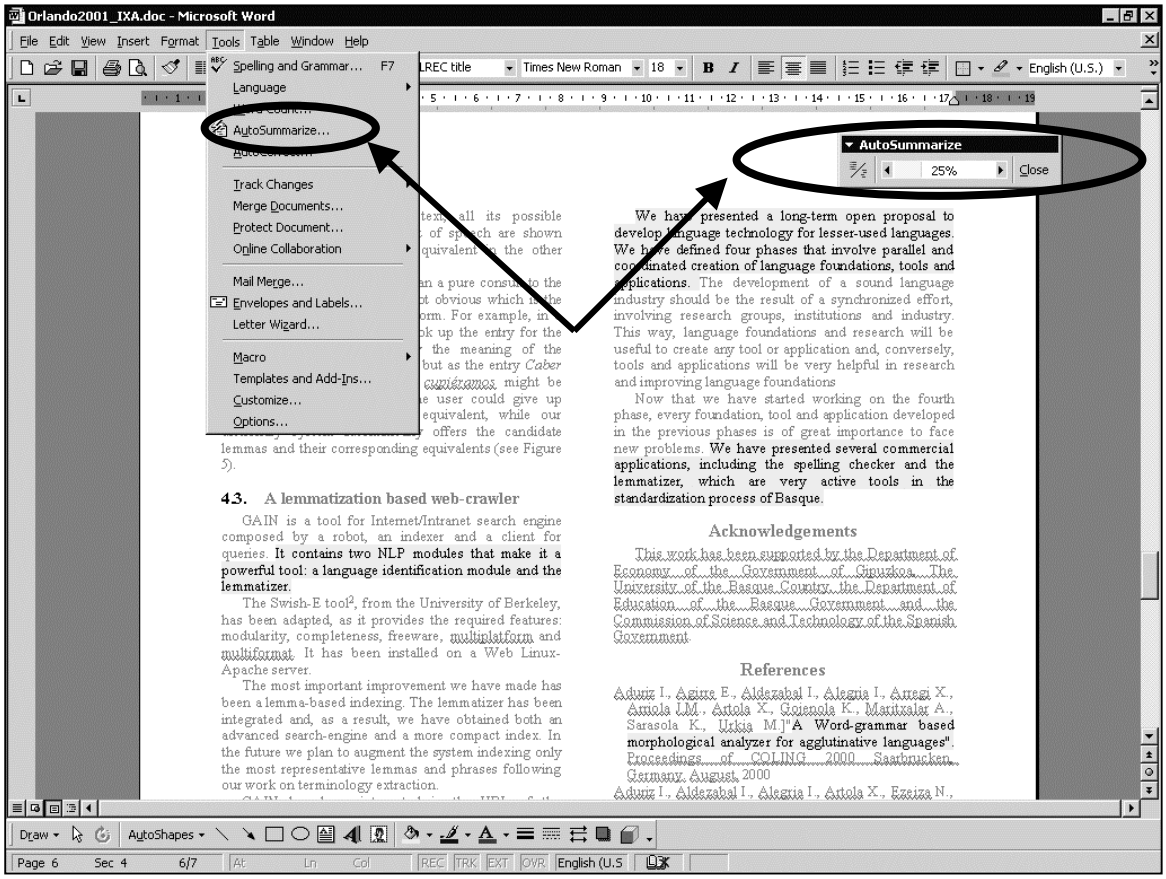
Atal honetan ere nazioartean hainbat sistema dago. Aipagarria da lehiaketak egiten direla nazioarteko laborategien artean ea nork lortzen dituen emaitzarik onenak; azkena MUC-7 izan da; 2001ean izan zen. Euskararen kasuan ere badira zenbait ekarpen. IXA taldean tresnak egin dira entitateen erauzketarako eta terminoen erauzketarako. Eleka-k *Xerlok* produktua garatu du, sistematikoki eta egunero hainbat

egunkaritako bertsio elektronikoetan enpresa edo entitate bati buruz egiten diren aipamenak bilatu ahal izateko.

6.2.3 Laburpen automatikoa (*Summarization*)

Dokumentuen laburpena automatikoki egitea bi eratara bidera daiteke. Modu errazena da testu zati edo esaldi esanguratsuenak hautatzea. Modu zaila erabiltzen denean, aldiz, ideia nagusiak detektatu, integratu eta testu berri bat sortzen da.

Testu-editore aurreratuek, hizkuntza-tresnen artean, eskaini ohi dute laburpenak egiteko aukera. Adibidez, 27. irudian ikus daiteke MSWord programak duen *Autosummarize* aukera. Hori eskatuta atzeko planoa horiz markatuta dutela azaltzen dira dokumentuko hainbat esaldi. Erabiltzaileak aukera dezake laburpenaren luzera, alegia, testu osoaren zenbateko portzentaia izan behar den laburpena.



27. irudia. Dokumentuaren laburpena MSWord programan

6.2.4 Dokumentu-sailkatzaileak

Sailkatze-sistemak oso baliagarriak dira dokumentu ugari kategoria multzo txiki baten arabera sailkatu behar izanez gero. Adibidez, hainbat albiste banatzea *kirola, nazioartekoa, kultura* eta *herrikoa* kategorien artean. Edota, Yahoo bezalako bilatzaile baten kasuan, web orri berri bat detektatzen duenean, zehaztea zein gaitan kokatu beharko den.

6.2.5 Dokumentuak bideratzea (*Routing*)

Aurrekoaren antzeko aplikazioa da hau, baina kasu honetan dokumentua sailkatzea bakarrik ez, dokumentuaren kategoriari dagokion helbidera edo sailera bidaltzen du dokumentua aplikazioak, dokumentuari postratamendu espezifiko bat emanez. Adibidez: enpresa batera egiten diren deiak, behar den sailera bideratzea. Edo egunkari batera heltzen diren berrien kasuan, ez da bakarrik zein saili dagokion baizik eta sail horretako erredaktore bati bidaltzea.

6.2.6 Dokumentuak multzokatzea (*Clustering*)

Aurreko aplikazioetan makina bat dokumentu aldez aurretik ezarritako kategoria multzo txiki baten arabera sailkatu behar ziren. Baina *clustering* egiten denean, aldez aurretik ez daude definituta kategoria posibleak. Abiapuntuan, hainbat dokumentu dauzkagu, eta bukaeran dokumentu horiek guztiak sailkatuta, haien arteko antzekotasunen arabera. Jakin beharko da geroago interpretatzen zergatik proposatu diren multzo horiek, zer adierazten duten azpimultzo horiek.

6.2.7 Dokumentuak iragaztea (*Filtering*)

Dokumentuen ezaugarri batzuk detektatu eta horren arabera dokumentu bera baztertu ala onartu egiten dute horrelako sistemek. Aplikazio honen adibide tipikoa posta elektronikoko *spam*-mezu guztiak detektatzea eta automatikoki alde batera uztea da.

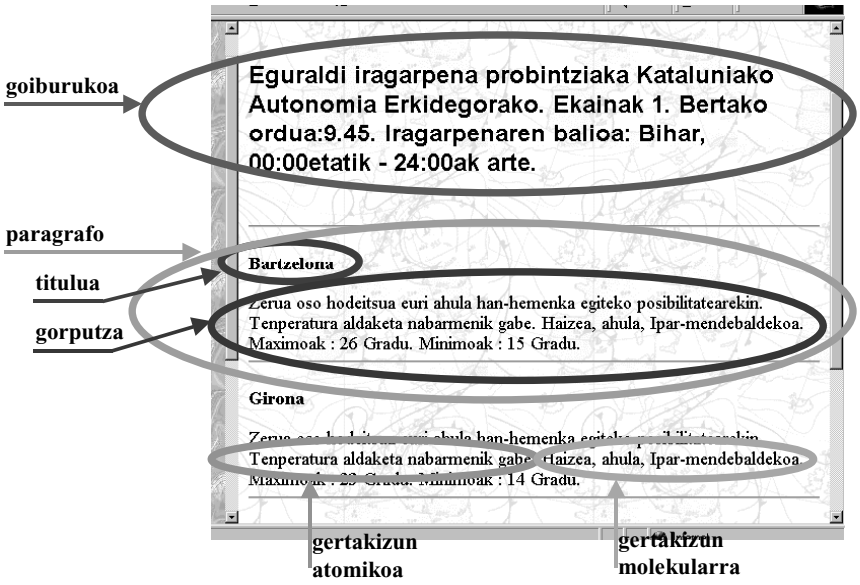
6.2.8 Testu-sorkuntza automatikoa

Testu-sorkuntza automatikoa informazio-erauzketaren kontrakoa da. Kasu honetan, ordenagailu barruan dauden datu konplexuetatik abiatuz (inprimakiak, datu kodetuak edo zenbakizko formatuan dauden informazioak...), datu horien edukia azalduko zaio erabiltzaileari bere hizkuntzan.

Adibidez, Multimeteo sistema euskaraz ere erabil daiteke, Parisko Lexiquest enpresa, UZEI eta IXA taldearen ikerketa-lanaren fruitu gisa. Sistema honek zortzi hizkuntzatan sortzen ditu eguraldi iragarpenak (espainiera, katalana, galiziera, ingelesa, frantsesa, alemana, nederlandera eta euskara), eta egunero irakur daitezke webeko helbide honetan: http://www.inm.es/wwi/MultiMeteo/Multimeteo.html

Egunero berritzen diren testu xume hauek automatikoki sortzen ditu programak. Baina programa horren betebeharra ez da itzultzea, testu-sorkuntza baizik. Alegia, zortzi hizkuntzetako bertsioak sortzeko abiapuntua ez da testu bat, hainbat zenbaki dituen matrize bat baizik (aurreikusten diren tenperaturak, haizearen norabidea, indarra...).

28. irudian ikus daiteke testu adibide bat, Bartzelonarako 2001eko ekainaren 2an sortu zena:



28. irudia. Multimeteo: hainbat datu abiapuntu gisa hartuta, euskarazko testua sortzen du

Jasotako datu numeriko guztiak eredu matematiko konplexuen bidez prozesatzen dira. Prozesu automatikoek simulatzen dute aldagai fisikoek hurrengo egunetan izango duten bilakaera, eta horrela, iragarpen meteorologikoetarako datu-matrizeak sortzen dituzte. Meteorologoak orduan aukera du datu-matrize horietan ukituak egiteko, alegia, bere eskarmentua erabiliz aurreikuspena osatu eta biribiltzeko. Ondorio gisa, 1. taulan ikusten dugun bezala, matrizeak hainbat ordutarako (3 orduko epeetan, INMren sistemaren kasuan) hurrengo datuak aurreikusten ditu: tenperatura (Te), haizearen norabidea (DD) eta indarra (FF), hodeiak, euria eta abar. Horrelako matrize bat lortzen da mapako puntu geografiko bakoitzerako.

hGMT	Nt	W	Wi	Wp	We	Wt	To	Top	Toe	Vv	Vp	Ve	Vt	Te	DTe	H	DD	FF	FFt	HN
03:00	3	2	1	2	2	2	0	0	0	0	0	0	0	18	3	0	1	1	0	9999
06:00	3	2	1	2	2	2	0	0	0	0	0	0	0	17	3	0	1	1	0	9999
09:00	3	2	1	2	2	2	0	0	0	0	0	0	0	21	2	0	1	1	0	9999
12:00	3	2	1	2	2	2	0	0	0	0	0	0	0	25	1	0	8	1	0	9999
15:00	3	2	1	2	2	2	0	0	0	0	0	0	0	25	2	0	8	1	0	9999
18:00	4	2	1	2	2	2	0	0	0	0	0	0	0	22	2	0	8	1	0	9999
21:00	4	2	1	2	2	2	0	0	0	0	0	0	0	18	2	0	0	0	0	9999
00:00	3	2	1	2	2	2	0	0	0	0	0	0	0	15	2	0	1	1	0	9999

6.3 Itzulpen automatikoa

Itzulpen-tresna zehatzetan sartu aurretik, itzulpen automatikoaren aplikazioei sarrera bat eskainiko diegu. Izan ere, itzulpengintza automatikoa izan zen amesturiko jomuga nagusia ordenagailua bera sortu zenetik.

6.3.1 Itzulpen automatikoaren garapen historikoa eta erronkak

Hizkuntzen arteko itzulpena betidaniko beharra izan da milaka urtetan. Itzulpengintza beste alor askori lotuta ikertu izan da, hala nola, linguistika, antropologia, psikologia, literatur teoria, filosofia, ikerketa kultural eta bestelako ezagutza-arloei, eta horrenbestez, ikuspegi teoriko ugarietatik aztertu izan da.

Zalantzarik ez da hizkuntza batetik bestera itzultzerakoan fenomeno asko izan behar direla kontuan, eta itzultzen den testuaren tipologiaren arabera, konplexuagoa edo sinpleagoa izan daitekeela itzulpen-prozesu hori. Tipologia horren barruan muturrekotzat jotzen dira literaturazko testuak eta testu teknikoak.

Itzulpengintza automatikoaren lehen pausuak ordenagailua bera sortu zeneko garai berean eman ziren. Baina urtetan zehar gorabehera handiak jaso ditu, hasierako ahaleginak zeharo geldiarazi zituen 1964an ALPAC txosten ezagunak. 1970. urteaz geroztik, berriro hartu zuen indarra adimen artifizialeko teknika berrien eskutik (gramatika-erregeletan oinarritutako sistemak), METEO sistemaren emaitza onak zabaldu nahian. 1995. urteaz geroztik, corpus elebidun zabalen agerpenarekin batera hirugarren aroa zabaldu da itzulpengintza automatikoan.

Baina erabateko automatiko den itzulpena egiteko ageri diren mugak direla-eta, itzulpenetan lagungarri izateko sistemak garatzea izan da itzulpen automatikoan indartu den atal garrantzitsua. Itzultzaile profesionalek eta itzulpen-agentziek itzulpen automatiko hutsa baino **itzulpen-laguntzak** nahiago omen dituzte. Itzultzailearentzat bereziki prestatutako **lan-estazioak** (translator *workstation*) direnak, alegia, itzulpen-memoriaz gain beste laguntzak eta funtzioak ere dituztenak:

- Hiztegi elektronikoen kontsulta aurreratua.
- Terminologiaren erauzketa, bilaketa eta kudeaketa.
- Testu-edizio eleanitza errazten duen hainbat leiho era koordinatuan kudeatzen dituen interfazea.
- Softwarearen lokalizazio-lanak errazteko tresnak.
- Bitestuak (corpus elebidunak) parekatzeko tresnak.
- Europako hizkuntza guztietarako eta Asiako hainbatetarako baliagarritasuna.
- Beste hizkuntzetarako testu-ediziorako laguntzak.
- Eta nahi izanez gero, eta nahi den zatian, itzulpen automatiko osoa erabiltzeko aukera.

Giza itzulpenean laguntzeko baliabide horiek gizakiaren eta ordenagailuaren arteko elkarrekintzako hiru estadiotan erabiltzen dira:

- **Aurre-edizioa.** Sorburu-testuaren prestaketa lantzen da fase honetan, besteak beste, testua erabilerraztu, aurreikus daitezkeen arazoak markatu eta horiek ebazpidean jartzen dira. Helburu horiek erdiesteko jarduera automatiko ohikoenak hauexek dira:
  - Terminoen edo hitz berezituak identifikatu.
  - Testua puskatan antolatu.
  - Aurre-itzulpena edo itzulpen-aukera bakarra duten sorburu-hitzak automatikoki itzuli.
- **Post-edizioa.** Xede-testuaren bertsio bat lortutakoan abiatzen da post-edizioa. Fase honetako eginkizunik aipagarrienek egiaztapenarekin eta zuzenketarekin dute zerikusia. Ortografia, sintaxia, estiloa eta hitzak itzultzeko beste adiera posibleak jorratzen dira zuzenketa-saioetan.
- **Edizioa.** Itzultzailea bete-betean itzultzen ari dela gertatzen da edizioko elkarrekintza. Etapa honetan testua prozesatzeko aukerak eta informazio-kontsultak dira gehien ustiatzen direnak.

Aurretik itzulita dauden testuak datu-base batean gordetzeko aukera ematen dute **itzulpen-memoriek.** Itzulpen-memoriak itzultzeko testua memorian dauden adibideekin parekatzen du, itzuli beharreko testu zatiaren berdina edo antzekoa bilatu eta itzultzaileari proposatuz. Hau da, esaldi berri baten itzulpenari ekiterakoan, sistemak bilaketa azkarra egiten du memorian, eta bertan esaldi hori edo antzekorik lehendik itzulita dagoen detektatzen du. Aurkitzen duen edo dituen testu zatiak proposatu, eta ondoren itzultzaileak hautatzen du edo osatzen du itzultzen ari den testuaren itzulpen egokia.

Itzulpen-memorietan funtsezkoa da corpus parekatuen kudeaketa (bitestuak esaten zaie). Zer sartu eta zer ez? Bakarrik kalitatezko itzulpenak? Edo testu asko nahiz eta kalitatea beti oso ona ez izan? Beste arazo bat itzulpen-unitatea zein den zehaztea da: perpausa? izen-sintagma? Euskara hizkuntza eranskaria izanik, kasu-atzizkiak bereizten eta tratatzen jakingo duten funtzionalitate bereziak beharko dira horrelakoak tratatzeko.

Itzulpen-memoriak oinarritzat dituzten hainbat sistema daude, eta batzuek beste batzuek baino aukera gehiago ematen dituzte. Ezagunen artean hauek ditugu: Wordfast, Trados-en Translation Workbench, Atril-en Déjà Vu, IBMren TranslationManager, Eurolang-en Optimizer eta STARen Transit.

Azken hamar urteotan, alde batetik, itzulpen-memoria horien erabilera ikaragarri zabaldu da, eta bestetik, ekarpen esanguratsuak egin dira **corpusaren gainean metodo estatistikoak** erabiliz (testuen analisia, bitestuen parekaketa....).

Etorkizunari begira erronka nagusietako bat **hiru metodologia horiek** (erregelen erabileran oinarritzen den metodologia klasikoa, itzulpen-memoriak eta estatistikoa) **integratzean** datza. Hainbat ikertaldek etorkizuneko erronka gisa ekin dio lexikoa aberasteko arazo kritikoa gainditzeari, horretarako azkenaldian bildu diren corpus eta baliabide lexikal erraldoiak erabiliz. Bestalde, interes handia dago **teknika berriak** esperimentatzeko: sare neuronalak, prozesaketa paraleloa, adibideetan oinarritutako itzulpen automatikoa.

Hizketaren tratamendua itzulpen-sistemetan integratzea da beste lerro berritzailea. bi proiektu aipagarri dabiltza gai horrekin: C-STAR eta Verbmobil. Japoniako ATR ikerzentroa, AEBko Carnegie-Mellon

Unibertsitatea eta Alemaniako Karlsruhe Unibertsitatea elkarlanean dabiltza (C-STAR partzuergoan) telefono bidezko itzulpen-sistema bat garatzen, ingelesa, japoniera eta alemanerako, edozein pertsonarentzat eta denbora errealean lan egingo duena, eta hasiera batean bi eremutan bakarrik aritzeko: hotel-erretserbak eta biltzarretan izen-emateak egiteko elkarrizketetan.

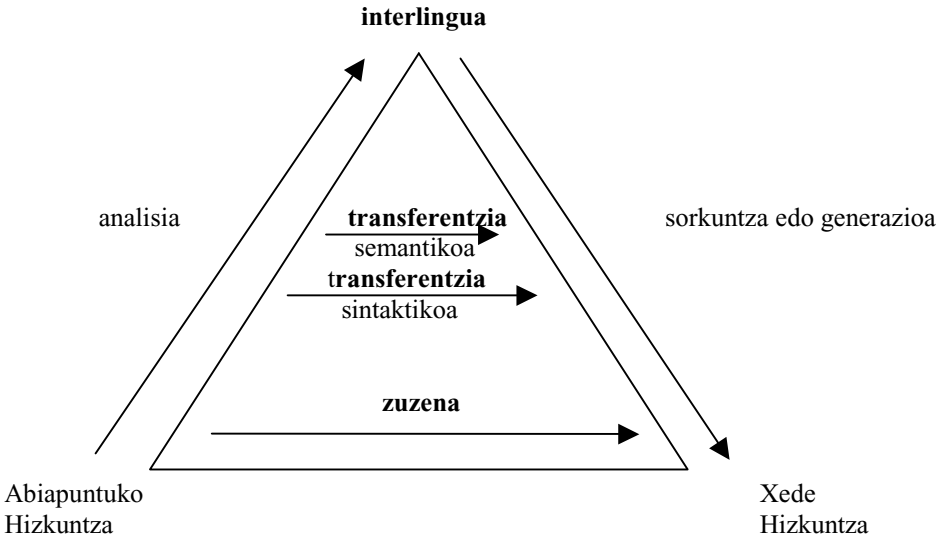
Alemaniako gobernuaren finantziazioarekin Verbmobil proiektu erraldoian enpresen arteko negoziaketetarako baliagarri litzatekeen hizketa-laguntza bati ekiten diote. Elkarrizketa, hizketa-ezagutza eta itzulpen automatikoa, hirurak batera, lantzen dituzte partaideek. Proiektuak begirada asko, arreta handia, bereganatu ditu, baina aditu gutxik espero du hortik aurrerapauso esanguratsuak epe laburrean etortzerik. Bitartean, merkatuan dabiltzan zenbait sistemak hizketako sarrera eta irteera onartzen dute, baina beti hiru modulu erabilita: hizketa-ezagutza testura, gero itzulpena testutik testura, eta bukaeran, testutik hizketarako bihurketa.

Internet-zerbitzuetako erabiltzaileak bilatzen duena informazioa da, edozein hizkuntzatan dagoela. Hemen itzulpena ez da xedea, tresna baizik. Erabiltzaileak nahi duena lortzeko behar dira bildu informazio-bilaketa, informazio-erauzketa, laburbiltze-sistemak eta itzulpen automatikoa; denak batera, baina modu gardenean batuta, eta erabiltzaileak konplexutasun horretaz konturatu gabe. Ildo horretatik doaz *"cross-lingual information retrieval",* laburbiltze-sistema eleanitzak, eta testu-sorkuntza eleanitza datu-baseetatik. Beraz, hemendik urte batzuetara, agian ez ditugu ikusiko itzulpen automatiko "hutseko" sistemarik, baizik eta ordenagailu bidezko **hainbat tresna eta aplikazio non itzulpen automatikoa osagaietako bat baino ez den izango.**

6.3.2 Itzulpen automatikorako estrategia klasikoak

Itzulpen automatikorako hurbilpen klasikoan hiru estrategia erabili ohi dira, baina badira bestelakoak ere. Estrategia horien arteko aldea, erabiltzen den informazio edo ezagutza linguistikoaren araberakoa dela esan daiteke. 29. irudiko Vauquois-en triangelua erabiliz azalduko ditugu.

Itzulpen zuzena. Abiapuntuko Hizkuntzatik (AH) zuzenean itzultzen da Xede Hizkuntzara (XH), analisi sintaktiko edo semantikorik egin gabe. Gehienetan hizkuntza biak oso gertu daude, eta nahikoa da hitzez hitz baliokideak lortzea. Batzuetan, Xede Hizkuntzako hitzen ordenari dagozkion egokitzapenak ere egiten dira, baina testuinguru oso lokala kontuan harturik. Aspaldiko sistemetan nahiz egungo itzulpen automatikoko softwaretan erabiltzen den estrategia da. Adibide gisa, katalana eta espainieraren arteko itzulpenak egiten dituen InterNostrum sistema dugu (www.InterNostrum.org). Sistema horrek bi erabilera eskaintzen ditu doan eta Internet bidez. Alde batetik, testu bat sartuta berehala itzultzen du itzulpenaren lehen zirriborro gisa erabili ahal izango dena; 30. irudiko adibidean, mezu elektroniko labur bateko hiru lerro itzuli dira. Beste erabilera posiblea web nabigatzailea da, itzultzailea *browser* batekin integratzen da eta web helbide bat emanda, horren testu guztia itzuli eta beste hizkuntzan aurkeztuko du; 31. irudiko adibidean katalanez bakarrik ikus daitekeen web orri bat aurkezten da, behin sistema espainierara itzuli eta gero.



29. irudia. Itzulpen automatikoaren hiru estrategiak: zuzena, transferentzia eta interlingua

Transferentziazko sistemak. Sistema hauek AHren eta XHren analisi gramatikalak egitea eskatzen dute, eta transferentzia, egitura sintaktikoan (zuhaitz sintaktikoak) nahiz semantikoan (adierazpen semantikoa) egiten da. Hau da, AHn itzuli behar den testu zatiaren analisi sintaktiko edo semantikoa egiten da, eta transferentzia egin eta gero XHren generazioa egiten da. Transferentzia gauzatzen den mailaren arabera, transferentzia-sistema desberdinak eraikiko dira. Oro har, zenbat eta abstraktuagoa izan transferentziazko errepresentazioa, orduan eta errazagoa da transferentzia-modulu egokia eraikitzea.

Interlingua. Sistema honetan AHko esaldia lengoaia neutral baten bitartez analizatzen da, eta lengoaia neutral hori da XHk erabiltzen duena esaldia sortzeko. Estrategia honek transferentziaren beharrik ez edukitzea ahalbidetzen du.

Vauquois-en triangeluak, bertikalean, analisi/generaziorako beharrezkoa den esfortzua ilustratzen digu, eta horizontalean transferentziarakoa. Puntu gorenean transferentziako esfortzua minimoa da, analisi eta generazioa, ordea, maximoa.

Jakina, triangelu honek hiru estrategiak erabat bereizita adierazten ditu, baina elkarren arteko konbinazioa posible da. Esaterako, sistema batek interlinguako eta transferentziazko elementuak erabil ditzake bere errepresentazioan. Edota transferentziako sistema batek ezaugarri sintaktiko, semantiko eta lexikalak erabil ditzake bere errepresentazioan. Badira zuzeneko estrategia eta transferentziazkoa konbinatzen dituztenak.

Transferentziazko nahiz interlinguako sistemek ezagutza linguistiko aberatsaren beharra eskatzen dute, hala nola, fonetika eta fonologia, morfologia, sintaxia, lexikoia eta semantika, eta pragmatika eta estilistika.

6.3.3 Itzulpen automatikorako produktuak

Produktu ugari dago merkatuan salgai edo erabilgarri testu-itzulpenean laguntza emateko, baina euskara tratatzen duen sistemarik ez dago oraindik. Hala ere, zenbait egin dira. Ikus ditzagun zein izan diren saio horiek munduan dauden produktu nagusiak aztertu baino lehenago.

IXA taldean Matxin sistema garatzen ari da. Euskararako itzulpen automatiko eleanitzerako prototipo bat eraiki da eta Adibideetan Oinarritutako Itzulpen Automatikoa baliatzeko proposamen bat egin da.

Opentrad proiektuan es-eu itzultzaile bat prestatzen ari da. 2006rako euskarak kode irekiko itzulpen automatikoko sistema izango du; testuak eta web orriak itzuliko ditu erregeletan oinarritutako hurbilpen klasiko bati jarraituz eta Egoera Finituko teknologiaren erabilera handia eginez. Hala ere, sistema honek ez du inoiz ere kontuan hartuko adiera-desanbiguazioa, hiztegiko lehenengo adiera hartuko du itzulpenerako beti. Lau unibertsitate eta hainbat enpresaren artean gauzatzen ari dira proiektua. Alacanteko Unibertsitateko Transducens taldea (Internostrum sistemaren egilea), Euskal Herriko Unibertsitateko IXA ikerketa-taldea, Vigoko Unibertsitateko Linguistika Informatikoko Mintegia eta Kataluniako Unibertsitate Politeknikoko TALP ikerketa-zentroa arduratu dira ikerketaz. Ikerketa-taldeek ateratako emaitzak produktu bihurtzeko lana, berriz, Elhuyar Fundazioaren, Bartzelonako Zientzia Parkeko Thera Hizkuntz eta Konputazio Zentroa enpresaren, Galiziako Imaxin Softwaree enpresaren eta Eleka Ingeniaritza Linguistikoaren esku dago.

Kataluniako *El Periódico* egunkaria egunero gaztelaniatik katalanera itzultzen laguntzen duen sistema eraiki duen ATS enpresak prototipo bat egin zuen gaztelaniatik euskarara testu ofizialak itzultzeko.

Itzulpen-memorien edo corpus lerrokatuen bilketan eta kudeaketan lan egiten dute, batetik, DELI taldeak eta, bestetik, EHUko Zientzia eta Teknikako Fakultateko Arantza Casillas-ek. Code&Syntax eta Eleka enpresetan lankidetzan ibili dira talde horiekin. Code&Syntax enpresak Tumatxa internet-zerbitzua atera du memoriak Internet bidez kontsultatzeko. EIZIE eta IVAP erakundeen web orrietatik itzulpen-memoriak jaitsi daitezke TMX formatu estandarrean.

Azkenik, zenbait azterketa egin dira metodo estatistikoen bidez Valentziako UPV unibertsitatean es-eu itzulpenak egiteko (González et al., 2004).

Euskararen inguruko produktuak alboan utzita, azter ditzagun merkatuan itzulpen automatikoaren barruan dauden produktuak. Adituek estimatzen dute 1.000 produktu-edo salgai direla merkatuan gaur egun (askoz gehiago, noski, hizkuntza-bikoteak bereiziz gero). Merkatuen dabiltzan produktuen zerrenda bat ikusi nahi izanez gero jo John Hutchins-ek Webean argitaratzen duen “Compendium of Translation Software directory of commercial machine translation systems and computer-aided translation support tools” ikustera (ourworld.compuserve.com/homepages/WJHutchins).

PCrako itzulpen-sistema gehienak AEBn eta Japonian sortu dira baina Europan badira produktu oso ezagunak: Comprendium eta T1 (Sail Labs), Personal Translator PT (Linguatec), iTranslator series (hasieran Lernout & Hauspie, gero Mendez), Reverso (Softissimo). Europako hizkuntza handientzat ere egin dira beste sistema batzuk, adibidez: PeTra (italiera eta ingelesa); Al-Nakil (arabiera, frantsea eta

ingelesa); Winger (daniera, frantsesa eta espainiera ingelesarekin); PARS (errusiera eta ukraniera ingelesarekin) edo TranSmart (finlandiera–ingelesa)

Sistema gehienek bertsio desberdinak eskaintzen dituzte enpresa handientzat, itzultzaile profesional independenteentzat eta noizbehinka mezu elektroniko bat edo web orri bat itzuli nahi duen etxeko erabiltzaileentzat. Oro har, itzulpenaren kalitatea ez da onargarria itzultzaile profesionalentzat, baina balekoa izan daiteke pertsona arruntentzat noizbehinkako lanak egiteko, esaterako, testu bat norberaren hizkuntzara itzultzerakoan lehenengo zirriborroa lortzeko, edo beste pertsona batzuekin beste hizkuntza batez komunikatzeko, posta elektronikoz adibidez.

Ordenagailu pertsonaletarako hainbat sistema merkaturatu dira web orriak eta mezu elektronikoak itzultzeko helburu soilarekin. Baina Internet dela eta, itzulpen-eskaera ikaragarri hazi da azken urteetan. Mementoan bertan nahi dugu itzulpena, eta ahal bada, on-line, beste programa edo konputagailu batera jo gabe, zuzenean Internet bidez (edo telefono mugikorraren bidez). Hizkuntza arrotz batean dagoen informazio bat ikusi eta berehala jakin nahi dugu zeri buruzkoa den, gutxi gorabehera bada ere. Ondorioz, gero eta gehiago erabiltzen dira **itzulpen automatikoko zerbitzuak Interneten**, eta gainera horietako asko doakoak dira, adibidez hauek:

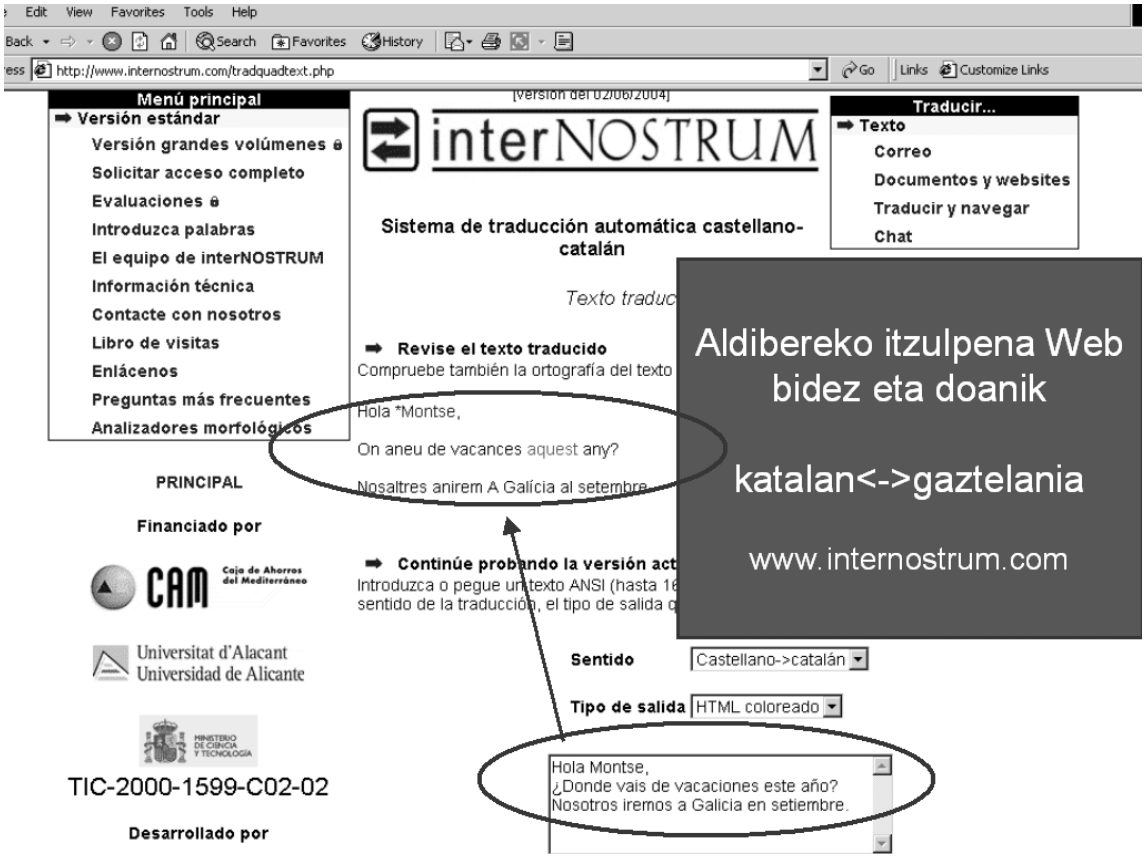
- Amikai (ingelesa, frantsesa, gaztelania, alemana, japoniera, txinera, portugesa, italiera eta koreera).
- ATS AutomaticTrans (gaztelania, katalana, portugesa).
- BabelFish (ingelesa, frantsesa, gaztelania, alemana, italiera, portugesa, eta koreera, txinera, japoniera). Mundu mailan sistemarik ezagunena hau da.
- Comprendium (katalana, gaztelania, frantsesa, ingelesa, alemana, errusiera).
- FreeTranslation (ingelesa, frantsesa, gaztelania, alemana...)
- interNOSTRUM (gaztelania, katalana).
- Reverso (ingelesa, alemana, frantsesa, gaztelania, errusiera)
- Systran (ingelesa, frantsesa, alemana, gaztelania, italiera, nederlandera, portugesa, errusiera).
- Worldlingo (ingelesa, frantsesa, gaztelania, alemana, portugesa, italiera, japoniera, errusiera, txinera).

Itzulpen automatikoaren erabiltzaile gisa Europako Batzordea da aipagarria, urtero itzultzen baititu 200.000 orri Systran sistema erabilita, non erabiltzaile tipikoa hizkuntzalaria ez den funtzionarioa den, dokumentu bateko informazioa aztertu nahi duena edo bere hizkuntza ez den testua itzultzeko lehenengo zirriborroa lortu nahi duena. Hala ere, Batzordearen Itzulpen Zerbitzuak bere lan-estazioa sortu du, EURAMIS, dauzkan hizkuntza-baliabideei probetxu handiagoa ateratzeko (batetik, Eurodicautom eta CELEX datu-baseak, eta bestetik, urteetan egindako itzulpen guztien bilduma).

6.3.4 Erabileraren zenbait adibide

Esan bezala, bi erabilera nagusitan aurkitzen ditugu itzulpen-sistemak:

Alde batetik, testu bat itzultzeko lehen zirriborroa sortzen dugu. Geroago, erabiltzaileak berak egokitu eta zuzendu beharko du zirriborroa, hau da, post-edizioa beharrezkoa izango da. Adibidez, katalana eta espainieraren arteko itzulpenak egiten dituen InterNostrum sistema dugu (www.InterNostrum.org). 30. irudiko adibidean mezu elektroniko labur bateko hiru lerro itzuli dira.

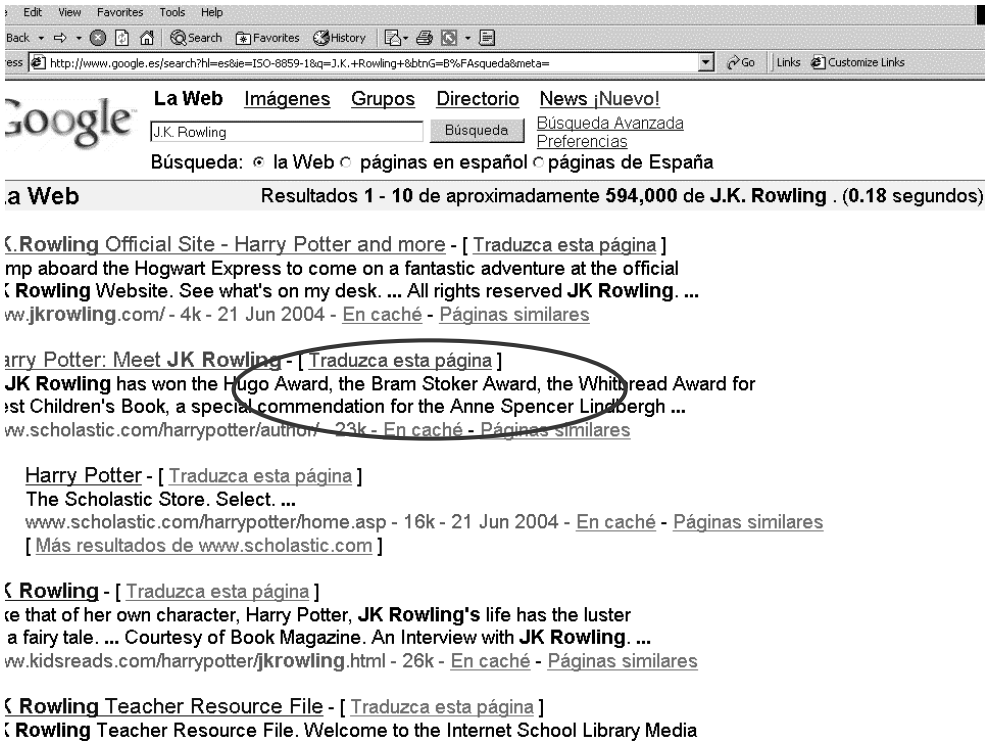


30. irudia. InterNostrum itzulpen-sistema

Beste erabilera, testu bat ulertu ahal izateko laguntza da. Adibidez, ulertzen ez dugun hizkuntza batean dauden web orriak itzultzen dizkigun web nabigatzailea. Kasu honetan itzultzailea *browser* batekin integratzen da, eta web helbide bat emanda, horren testu guztia itzuli eta beste hizkuntzan aurkezten digu. Bi adibide aurkezten ditugu, 30. irudian katalanez bakarrik ikus daitekeen web orri bat aurkezten da, behin InterNostrum sistemak itzulita.

31. irudia. Internostrum, Interneten nabigatzeko itzulpen-laguntza gisa

Beste adibidea Google bilatzailearena da, hainbatetan eskaintzen baitigu web orriaren itzulpena.

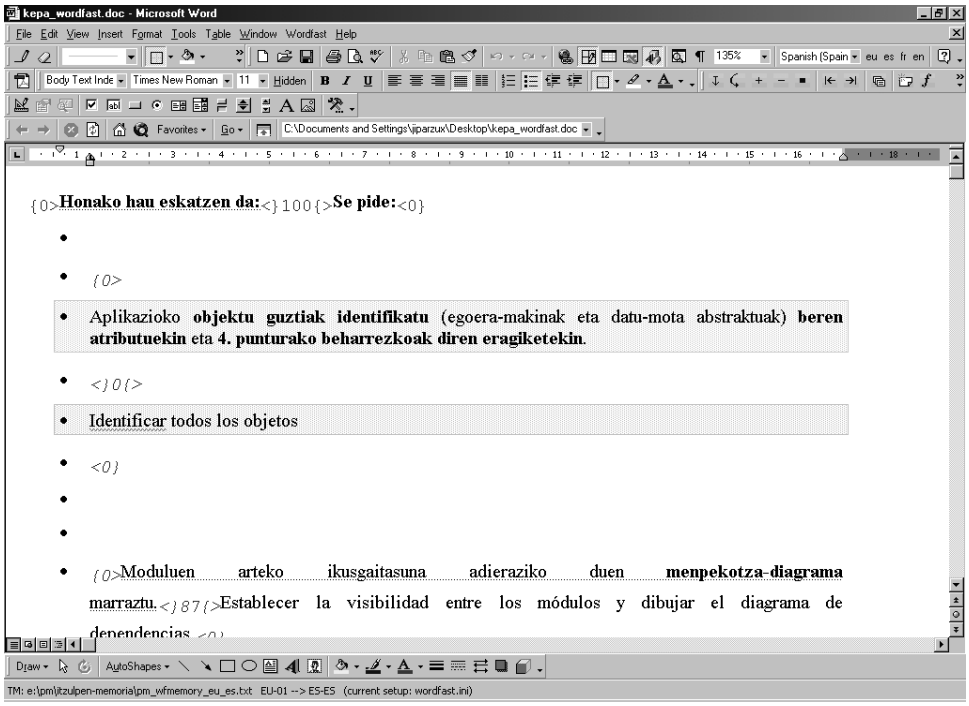


32. irudia. Google-k hainbatetan orrien itzulpenak eskaintzen ditu

33. irudian ikus daiteke Google-k gaztelaniara itzulitako web orri bat, Harry Potter eleberriaren egilearena, hain zuzen. Ikus daitekeenez, hitz guztiak ez dira itzuli: html kodean testu moduan azaltzen zena bai, irudi moduan azaltzen dena ez, ordea. Google-k erabiltzen duen programa itzultzailea Systran da. Ikus daitekeenez, itzulpena ez da perfektua (*"soy una persona extraordinario afortunada, haciendo lo que amo lo más mejor posible en el mundo" / "I am an extraordinarily lucky person, doing what I love best in the world"*), baina bai baliagarria testua zertaz ari den jakiteko eta ideia orokorra lortzeko. Hala ere, zenbait esaldi modu egokian itzuli dira (*"La recompensa más grande es el entusiasmo de los lectores" / "The greatest reward is the enthusiasm of the readers"*).

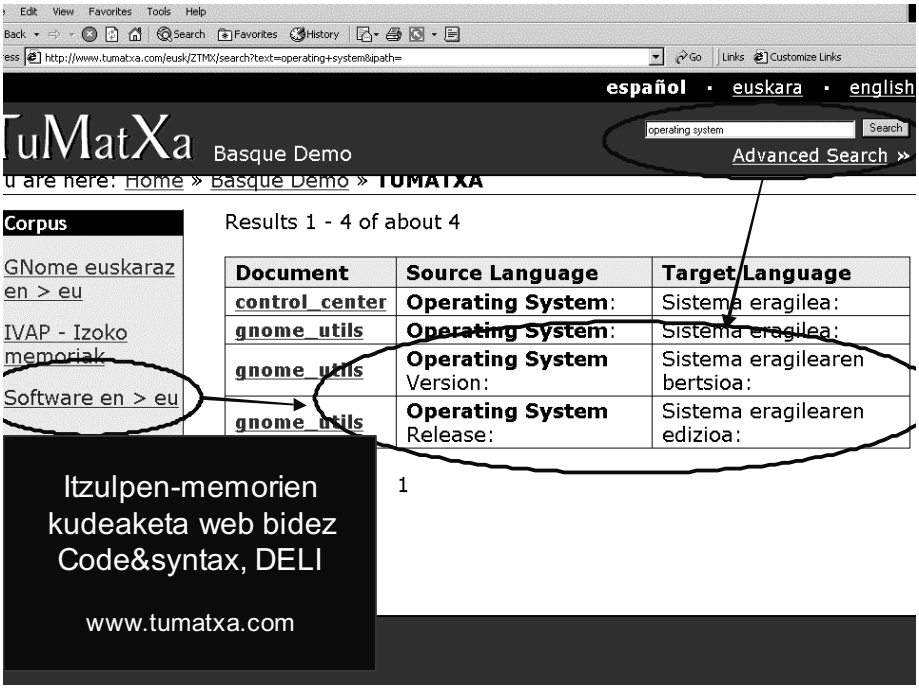
33. irudia. Google, orri baten itzulpena erakusten

- Aplikazio mota honetan, azken urteotan garrantzia handia hartu dute ***itzulpen-memori***ek.



34. irudia. Wordfast: itzulpen-memoriak erabiltzeko sistema

Azkenaldian itzulpen-memoriak Internet bidez kontsultatzeko sistemak ere plazaratu dira. Adibidez, 35. irudian erakusten den Tumatxa sistema.



35. irudia. Tumatxa: itzulpen-memoriak Internet bidez kontsultatzeko sistema

6.4 Interfazeak

Konputagailuaren erabileraren zabalkuntza areagotzeko, hizkuntza-teknologiaren helburu garrantzitsu bat gizakiaren eta makinaren arteko komunikazioa erraztea da. Hor bi sistema mota aurkitzen dira: komunikazioa testu idatziaren bidez bideratzen dutenak eta hizketaren bidez egiten dutenak.

Gizakiaren eta makinaren arteko elkarrekintzan laguntzeko sistemek gorabehera asko izan dituzte historian zehar. 1980ko hamarkadan zabaldu ziren, batik bat datu-baseak kontsultatzeko. Oro har, sistema horietan bi eratako funtzionalitateak biltzen ziren: alde batetik, datu-basea kontsultatzeko programa, eta bestetik, erabiltzailearekin komunikatzeko programa bera. 1990eko hamarkadan ez ziren gehiegi zabaldu, baina azken urteotan hizketaren tratamenduaren eta Interneten zabalkuntzaren inguruan emandako aurrerapausoak direla eta, berriro doa goraka horrelako sistemen garapena.

Sistema horien helburua informazioa eskatzeko mementoan erabiltzaileari askatasun handiagoa ematea eta haren galderak hobeto ulertzea da. Jakina, eskaeraren arabera erantzun egokiak ematea ere helburu garrantzitsua da eta oso lotuta dago informazio-bilaketaren eta informazio-erauzketaren atalekin.

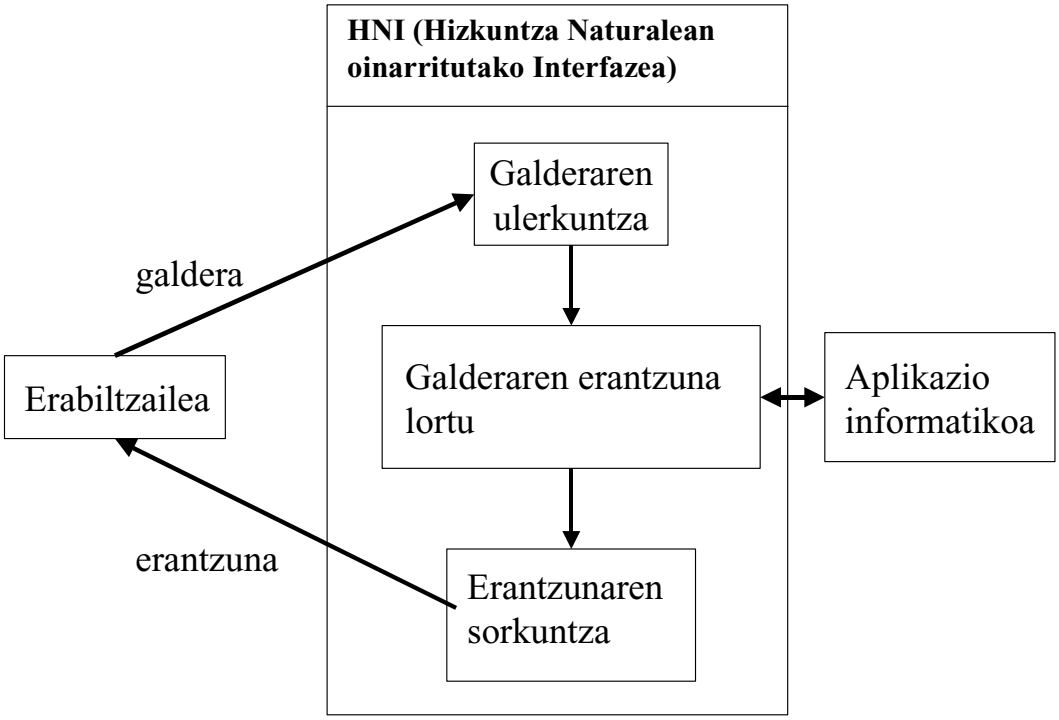
Oro har, eta laburbilduz, zeintzuk dira sistema hauen abantailak eta desabantailak?

- Erabiltzaileak ez du lengoaia artifizial bat ikasi behar. Horrela, noizbehinkako erabiltzailearentzat dira egokiak, baina komeni zaigu horren inguruan ñabardura bat egitea: interfaze hauek ulertzen duten hizkuntzaren estaldura murriztua da; beraz, hizkuntza osoaren azpimultzo bat tratatzeko gai dira, besterik ez; eta zaila da erabiltzaileari murriztapen horren berri ematea. Gerta daiteke erabiltzailea galduta ibiltzea, jakin ere ez dakielako zer hitz, zer esaldi edo zer esateko modu erabil ditzakeen, hau da, ez duelako ondo asmatzen hizkuntzaren zer azpimultzo erabili behar duen.
- Erabiltzaileak hainbat forma diskurtsibo eta hizkuntz egitura erabiltzeko aukera ditu kontsultetan edo aginduetan. Ondoko adibidean ikus daiteke horrelako ezaugarri bat: bigarren galderan elipsia erabiltzen da, guztiz antzekoa izango zen galdera oso-osorik ez errepikatzearren:

```
- Zein da Frantziako hiriburua?
- ...
- Eta Italiakoa?
- ...
```

6.4.1 Lengoaia naturalean oinarritutako interfazea duten sistemen arkitektura

36. irudian hizkuntza naturalean oinarritutako sistemen arkitektura aurkezten dugu. Erabiltzailearen eta sistemaren arteko komunikazioa interfazearen bitartez lortzen da, eta interfaze barruan hiru modulu izaten dira azpilan hauek egiteko: 1) galdera ulertzea, 2) galderaren erantzunak eman behar duen informazioa lortzea, eta 3) erantzuna sortzea.



36. irudia. Lengoaia naturalean oinarritutako interfazearen antolaketa

6.4.1.2 Galdera ulertzea

Makinak ulertu behar du erabiltzaileak zer eskatzen duen, gero bere agindua bete ahal izateko. Galdera zehatz-mehatz ulertu behar denez, modulu honetan ezin dira erabili akatsak eman ditzaketen teknikak. Gainera, sistemak galdera bat ondo ulertzeko, erabiltzaileari galde diezaioke zalantzak edo ondo ulertu ez dituen aspektuak argitzeko eskatuz.

Galderak ulertzeko, zenbait prozesu linguistiko aplikatu behar dira: lexikala, morfologikoa, sintaktikoa, semantikoa eta pragmatikoa. Maila ilokutiboa ere inportantea da; horretan erabiltzailearen asmoak edo helburuak identifikatu beharko dira.

Maila bakoitzak estaldura desberdina izango du. Horietako fase batzuk enoratu daitezke sistema batean. Batzuetan fase horiek sekuentzialki aplikatzen dira (maila bakoitzean sortzen diren emaitzak hurrengo mailarako sarrera izango lirateke), eta, beste batzuetan, elkarrekintza egoten da beraien artean. Sintaxia eta semantika tratatzeko badira hainbat teknika, liburu honetan bertan jada aipatu direnak.

Zeintzuk dira arazo zailenak? Hemen batzuk baino ez ditugu aipatuko:

- Preposizio-sintagma bakoitza zeinen modifikatzailea den zehaztea. Nahiko anbiguoa izaten da kontu hau (gogoratu *I see a man in the park with the telescope* adibidea).
- Kuantifikatzaileen esparrua (guztiak, zenbait...).

- Koordinazioa eta mendekotasuna.
- Anafora eta elipsia.
- Espresio ez-gramatikalak.

6.4.1.3 Galderaren erantzuna lortzea

Modulu honen bitartez aplikazio konkretuan dugun sistema informatikoarekin komunikatzen gara. Modulu hau aplikazioaren mende dagoenez, ez da oso garraiagarria, alegia, ezin da erabili beste gai baten inguruan egingo diren interfazeetan.

Modulu honetan eginbehar nagusienak hauek dira:

- Ulerkuntza modulutik behar den informazio guztia lortu dela konprobatu.
- Aplikazio informatikoa atzitu (adibidez, datu-base batekin egiten bada lana, SQL galdera bat sortu beharko litzateke).

6.4.1.4 Erantzuna sortzea

Modulu honetan erantzuna prestatzen da. Erantzun hau hainbat modutan eman daiteke: irudia, taulak, hizketa edo testua. Hizketa edo testu bidez ematen bada, testua bera sortu behar da. Testuak sortzeko tekniketan bi dira erabili behar diren aspektu aipagarrienak: *zer esan?* eta *nola esan?* Normalean bi modulu bereizten dira:

- Modulu estrategikoa: bere helburua ZER ESAN erabakitzea da. Horretarako, modulu honetan testua antolatzen da: zer terminologia erabili, zenbat informazio eman eta norainoko zehaztasunak emango diren.
- Modulu taktikoa edo osagai linguistikoa: lengoaia naturalaren sorkuntzaz arduratzen da. Konkretuki, hauexek lirateke erabaki beharreko gauzak: erantzunaren egitura sintaktikoa, osagai sintaktikoen ordena, zein hitz erabili kontzeptu bakoitza adierazteko (lexikoa)...

6.4.2 Bilakaera historikoa

ELIZA sistema izan zen lehenengoetariko bat. ELIZA programak psikologoaren papera hartzen du eta psikologoaren eta pazientearen arteko elkarrizketa bat simulatzen du. Sistemaren erantzunak aldez aurretik prestatuta daude. Galdera bat jasota, aztertzen du ea galderan agertzen diren hitz gako batzuk. Horien arabera aukeratzen du zein izango den erantzunaren patroia, eta galderako zenbait hitzekin osatzen du. Elkarrizketa txiki baten bidez jabetu gaitezke sistemaren mugez.

Elkarrizketa (E: Eliza; P: Pacient)**:**

P - *I have been very depressed lately*
E - *You say that you have been very depressed lately*
P - *Yes, my mother beats me and my father calls me names*
E - *Tell me more about your family*
P - *My father calls me nasty names*
E - *Who else in your family calls you nasty names?*
P - *I think the dog would if he could*
E - *Do you think it's likely that he could?*
P - *No, do you?*

E - *You don't seem very talkative today*
P - *I guess what is really bothering me is that I don't have any money*
E - *Can you elaborate on that?*
P - *The problem is that I know psychiatrists are expensive and I may not be able* to p*ay you*
E - *Please, continue, this may be interesting*
P - *I think you might think so*
E - *Perhaps in your fantasy we thought each other*
P - *Right doc*
E - *Earlier you said your mother beating you*
E - *Have you any other problems you wish to discuss*
P - ...

Sistema honek, beste batzuek bezala, hizkuntzaren tratamendurako oso teknika xumeak erabiltzen ditu; ez du egiten aurretik analisi sintaktikorik, edo morfologikorik. Esaldiak oso azaletik ulertzen ditu. Besterik gabe, esaldia identifikatzeko nahikoa den hitz gako bat edo batzuk bilatzen dira: patroiak. Patroi bat elementu-sekuentzia da, non elementu bakoitza hauetako bat izan daitekeen:

1. Hitza.
2. Edozein hitzekin bat etorriko den ikurra (+).
3. Edozein hitz-sekuentziarekin bat etorriko den ikurra (*).
4. <kategoria>. Kategoria baten izena, kategoria horretako edozein hitzekin bat etorriko da.
5. Aldagaiak. Edozein hitz edo hitz-sekuentziarekin bat etorriko direnak. Aldagaiak balio bat hartuko du patroia eta esaldia parekatzerik badago, eta geroago erabili ahal izango da balio hori erantzun bat antolatzeko.

Hauek izan zitezkeen sistema honen patroi batzuk:

?x	--> Can you elaborate on that?
?x	--> Please, continue, this may be interesting
I **?x**	--> You say that **?x**
?x mother ?y	--> Tell me more about your family
My mother **?y**	--> Who else in your family **?y**
?x father **?y**	--> Tell me more about your family
My father **?y**	--> Who else in your family **?y**

Analizatu behar den esaldia definitu diren patroiekin konparatuko da. Patroi bat esaldiarekin ondo parekatzen denean, patroi horri lotuta dagoen tratamendua aplikatuko da emaitza lortzeko. Baina askotan patroi bat baino gehiago dira egokiak esaldi baterako. Arazo hori konpontzearren, patroien multzo osoa modu hierarkikoan antolatzen da, patroi orokorrenetik patroi espezifikoenetara, eta horrela, patroi bat baino gehiago egokiak badira, hierarkiako espezifikoena aukeratuko da.

Patroi bat ere aktibatzerik ez dagoenean, tratamendu alternatiboak sortzen dira, adibidez: "Mesedez, azalduko didazu hori?".

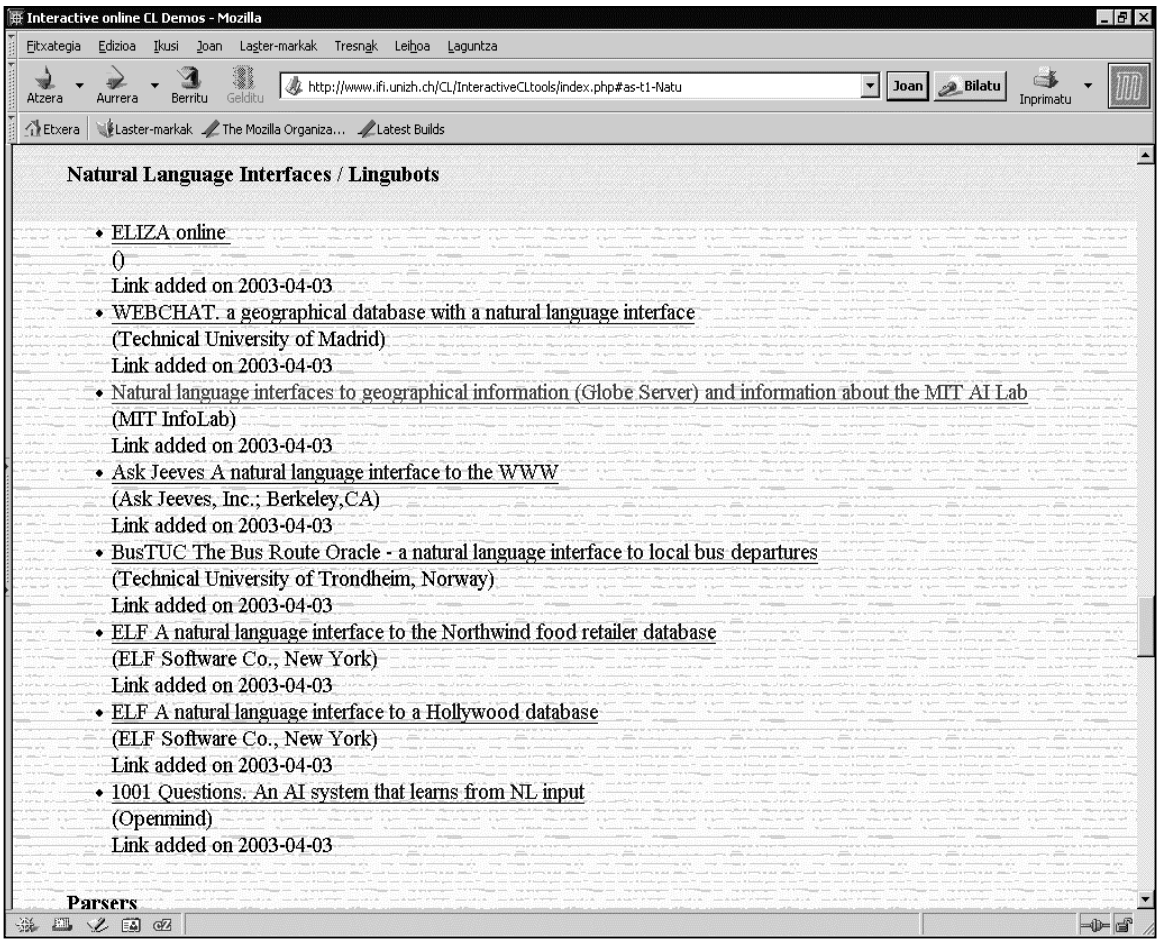
Patroi-parekatzea oso arrakastatsua izan zen LNPko hasierako sistema batzuetan. ELIZA sistemak (Weizenbaum, 1966) psikiatra batekin elkarrizketa bat simulatzen zuen; PARRY sistemak (Colby, 1975) paranoiko baten jokaera simulatzen zuen. LUNAR (Woods, 1972) izan zen analisi sintaktikoa egiten zuen

lehenengo sistema; analisi sintaktikoa trantsizio-sareen bitartez (ATN, *Augmented Transition Networks*) egiten zuen.

Datu-baseak kontsultatzeko beste sistema batzuk sortu ziren 80ko hamarkadan: LADDER, RENDEZVOUS, CHAT-80, eta abar. Barbara Grosz-ek garatutako TEAM sistema aipatzekoa da. Kontuan hartu behar da sistema hori komunikazioaren inguruko bestelako arazoak aztertzen hasi zela; konkretuki, helburu jakin batekin kolaboratzen duten bi agenteren (makina eta pertsona) arteko komunikazio mota aztertu zuen.

Geroztik datu-baseen kontsultarako erabiltzeaz gain, sistema hauek beste aplikazioetarako zabaldu ziren: sistema eragileak, sistema tutoreak, sistema adimendunak… Garai horretako adibideak hauek dira: LDC, TELI eta XCALIBUR. Horietako sistema batzuetatik produktu komertzialak ere sortu ziren: INTELLECT, HAL, Q & A.

90eko hamarkadan, sistema multimodalak izan ziren teknologia berrienak ekarri zituztenak; beraz, lengoaia naturalaren prozesamenduko teknikak interfaze multimodaletan integratzen dira. Aipatzekoa da ALFRESCO, pintura italiarrari buruzko datu-base batekin lan egitekoa. Galderetan posible zen erabiltzea irudietan ikusten ziren hainbat erreferentzia.



37. irudia. Lengoaia naturalezko interfazeak. Interactive online CL Demos

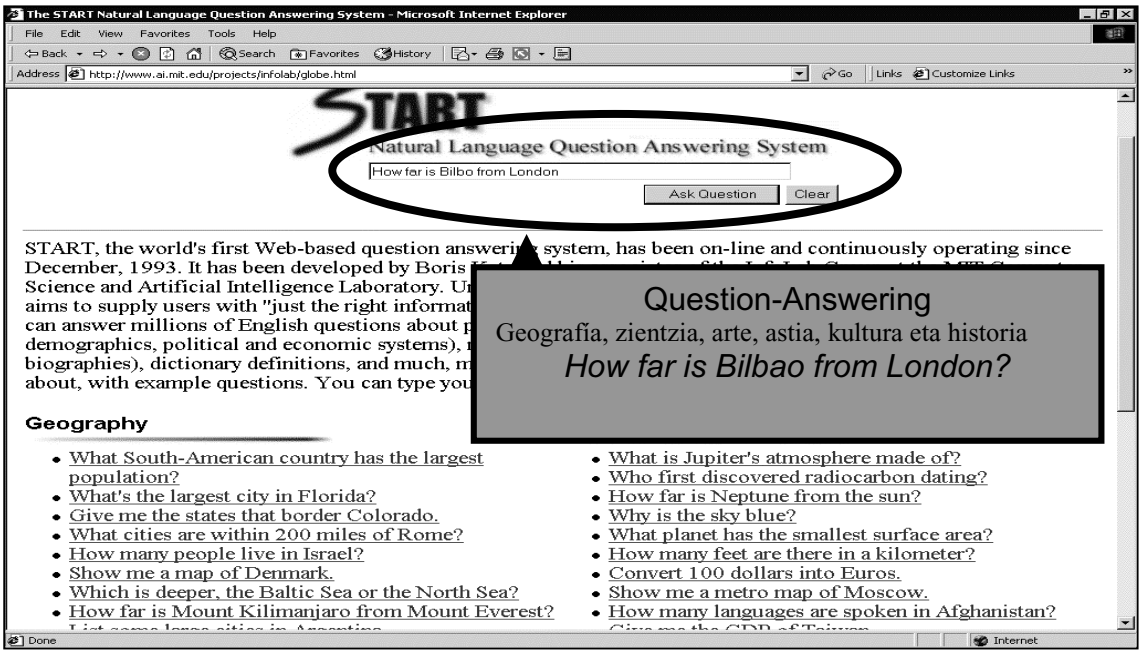
Hizketaren tratamendua integratzeak hobekuntza oso nabarmenak ekarri ditu azkenaldi honetan; horrela, testuaren tratamendua menu, hizketa, grafiko eta keinuen tratamenduarekin osatu da gizakiaren eta makinaren arteko elkarrekintza naturalago eta eraginkorrago bihurtzeko.

Helburu orokorreko sistemarik ez da luzaroan salgai egongo, baina badira dagoeneko aplikazio konkretuei lotuta dauden batzuk. Datu-baseetarako galdeketa-sistema ugari dago, batez ere ingelesez. Interneten *Interactive online CL Demos* izenburuko orrian (www.ifi.unizh.ch/CL/InteractiveTools.html helbidean eta *EN-Natural Language Interfaces / Lingubots* atalean) Eliza eta beste zazpi sistema hauek ikus daitezke martxan (ikus 37. irudia).

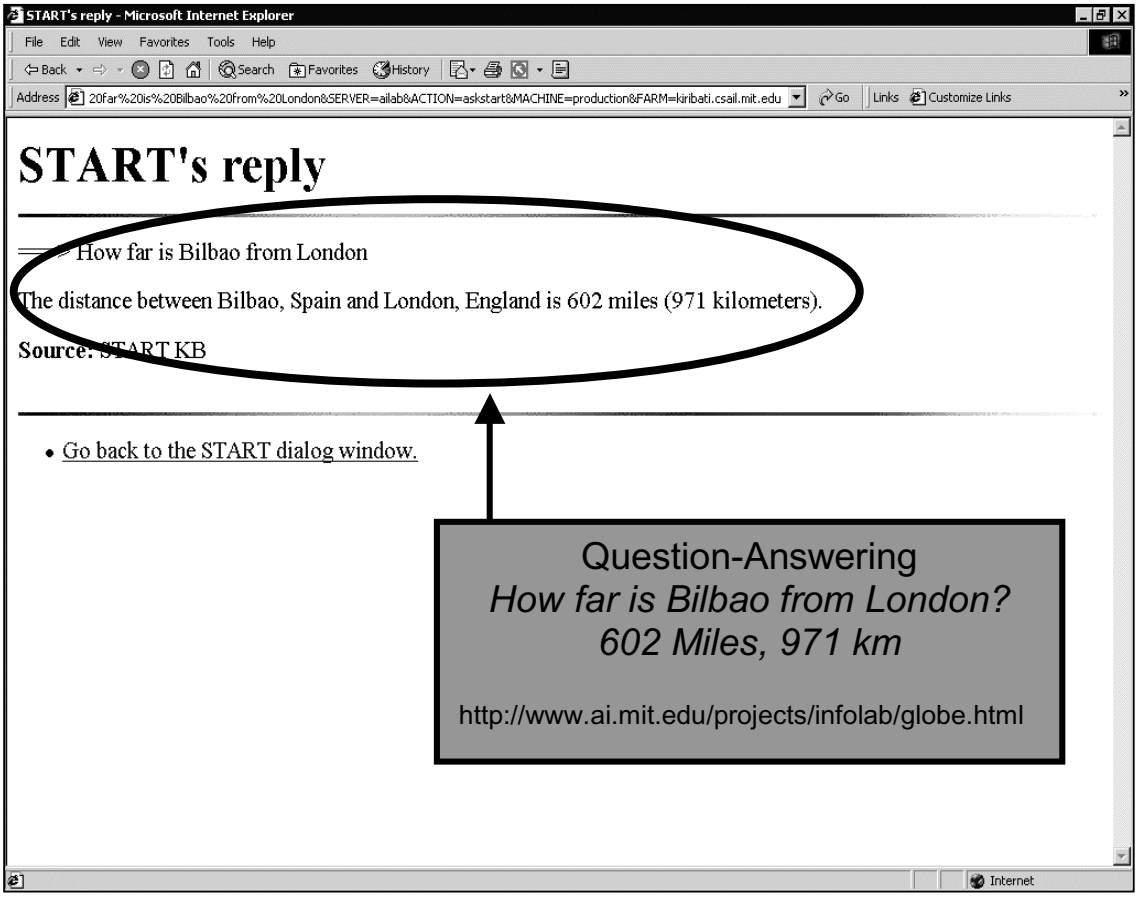
6.4.3 Galdera-erantzuneko sistemak (*Question Answering*)

Interfaze berezi hauen helburua da erabiltzailearen galderentzako erantzunak aurkitu eta itzultzea. Galderak lengoaia naturalez egiten dira eta erantzunak ere lengoaia naturalez eman behar dira. Erantzunaren edukia bilatzeko hainbat aukera izaten dira horrelako sistemetan: datu-base arrunt batean edo datu-base dokumental batean kontsultatuz, edota, besterik gabe, zuzenean Internetera joz (Internet osoa ala bere azpimultzo bat bakarrik kontuan hartuz)

Adibide gisa, MIT ikerketa-zentroan sortutako Start sistema (www.ai.mit.edu/projects/infolab/globe.html) aurkezten da 38. irudian. Sistemak gai hauei buruzko galderak onartzen ditu: geografia, zientzia, arte, aisia, kultura eta historia. Galderak egiteko zenbait adibide aurkezten ditu lehen orrian erabiltzailea galduegi ez ibiltzeko. 38. eta 39. irudietan ikus daiteke nola egin den galdera bat (*How far is Bilbao from London?*) eta sistemaren erantzuna (... *971 km*).



38. irudia. START: galdera-erantzuneko sistema



39. irudia. START galdera-erantzunak egiteko sistemaren erantzuna

6.5 Hizketaren tratamendua

6.5.1 Sarrera

Hizketako hitzak edo esaldiak ulertzea zaila da, hizkuntza idatzia ulertzeko arazoei ahozko hizkuntzaren problematika eransten zaielako: hitzen arteko mugak ez direlako guztiz markatzen hitz egitean, esaldien hasieran eta bukaeran erdialdean baino intentsitate txikiagoz ematen direlako, eta gainera, audio-seinale fisikoarekin batera inguruko zaratak ere sartu ohi direlako.

Sistema sendoenek, edozein erabiltzailerentzat eta edozein ingurutan erabiltzeko asmatu izan direnek, oso hitz gutxi ezagutzen dute, horien artean beti daudela zenbakiak. Adibidez, pertsona batek makina bati edo konputagailu-programa bati aginduak emango dizkio; agindu posibleen artetik bat hautatuko du zenbaki bat (edo hitz gakoren bat) ahoskatuz. Merkatu handia zabaldu da horrelako sistemak telefono bidezko zerbitzuetan integratzeko: aurretiko hitzordua, produktu-eskaerak, telefonogune-zerbitzuak ("zentralita-zerbitzuak"), e.a. Beste alde batetik, hizketaren ezagutzarik izan gabe ere, gero eta arruntago bihurtzen ari zaigu makinen ahots sintetizatuak entzutea: gasolindegietan, tabako-edariak saltzen dituzten makinetan, edo posta elektronikoko mezuak telefonoaren bidez irakurtzen dizkiguten sistemak. Natural Vox enpresa arabarrak aurretiko hitzordua (medikuarenean edo errenta-aitorpena egiteko) automatikoki lortzeko sistema telefonikoak ezarri ditu azken urteetan eta arrakasta handia izan du.

Ahozko hizkuntzaren tratamenduko zenbait teknika antzeko beste aplikazio batzuetan ere erabiltzen dira, esate baterako, eskuz idatzitako testuak irakurtzen dituzten sistemetan edota testu mekanografiatuen bertsio elektronikoa lortzen duten OCR programetan (*Optical Character Recognizer,* karaktere-ezagutzaile optikoak).

6.5.2 Historia

Arlo honen hasiera 1960ko hamarkadaren inguruan koka daiteke, hau da, sistema informatikoen bidez seinale akustikoak digitalki tratatzea posible izan zenetik. Ordenagailu bidezko hizketaren sorkuntza ere hamarkada hartan hasi zen. Fonema bakoitzaren ezaugarri akustikoak definitzen dituzten parametro fonetikoak definitu ziren orduan, eta horrela, hizketaren analisirako eta sintesirako teknika digitalak erabiltzeko bidea ireki zioten ordenagailuari. Teknika horien bidez ahotsa osatzen duten seinale akustikoak manipulatu ahalko ziren.

1970eko hamarkadan testua hizketa bihurtzen duten sistemak garatu ziren. Sistema horiei esker ozenki irakur daiteke ordenagailuan gorderiko testu bat. Aldi berean, hitz isolatuak ezagutzen zituzten sistemak sortu ziren, hizketaren analisiaren bidez hiztunaren nortasuna egiaztatzen duten sistemak eta hizketa prozesatzeko teknika berriak, hartara prozesamendu informatikoa hobetuz.

1980ko hamarkadan, aldiz, hizketa jarraituaren ezagutza-tekniketan egin ziren lehenengo aurrerapenak. Horrela, bada, ordenagailu batek gure enuntziatua ezagutzea nahi dugunean ez dugu hitzen artean etenik egin behar.

1990eko hamarkadan, hizkuntza desberdinetarako sintesia egiten duten produktu komertzialak egin ziren, aurreko garaietan baino naturalagoak eta malguagoak zirenak. Analisiaren arloan, hiztegi oso handiak integratuta edozein hitz ezagutzea lortu zen, eta geroago ondorio gisa, diktaketa automatikorako sistema komertzialak ere sortu ziren. Hamarkada hartan ere, elkarrizketa-sistemen hedapena suertatu zen. Elkarrizketa-sistema horiei esker hainbat jarduera egin ahal izan dira: txartelak erreserbatu; garraioen ordutegiei buruzko edota gainerako hiritarrentzako zerbitzuei buruzko informazioa eskuratu elkarrizketa baten bidez sistema informatiko batekin, esaterako telefonoz...

2000ko hamarkadan ohitu egin gara hizketa automatikoki tratatzen duten hainbat sistemarekin, batez ere telefono-deiak automatikoki bideratzen dituzten zerbitzuekin. Teknika aldetik, hizketa sintetizatuaren naturaltasuna lortze aldera jotzen da. Horretarako, batetik, sistema saiatzen da aurregrabazioetatik (corpusetik) hitz osoak bere osotasunean hartzen, hitz osoa fonemen bidez artifizialki sortu gabe. Eta bestetik, saiatzen da emozioa integratzen: tristura, poza... Ildo horretatik *abatarrak* erabiltzen dira: apllikazio informatiko batek aurpegi baten bitartez hitz egiten digu, keinuak eginez eta emozioak adieraziz.

6.5.3 Helburuak

Hizketaren tratamenduan oinarritzen diren aplikazioen helburu nagusia pertsonen eta sistema informatikoen arteko elkarrekintza ahotsaren bidez gauzatzea da. Hartara, gizakiak modu eraginkorrago batean baliatu ahal izango ditu sistema informatikoak, horiek inposatzen dituzten murriztapenak edota mugak gaindituz. Hori dela eta, lehendabizi ahotsaren bidez ordenagailu batekin komunikatu ahal izateak dituen abantailak eta zenbait muga aipatzen dira. Horrez gain, sistema multimodalak ere lantzen dira, hau da, ahotsaz gain testua, irudia edota keinuak ere integratzen dituzten sistemak egiten dira.

6.5.4 Atalak

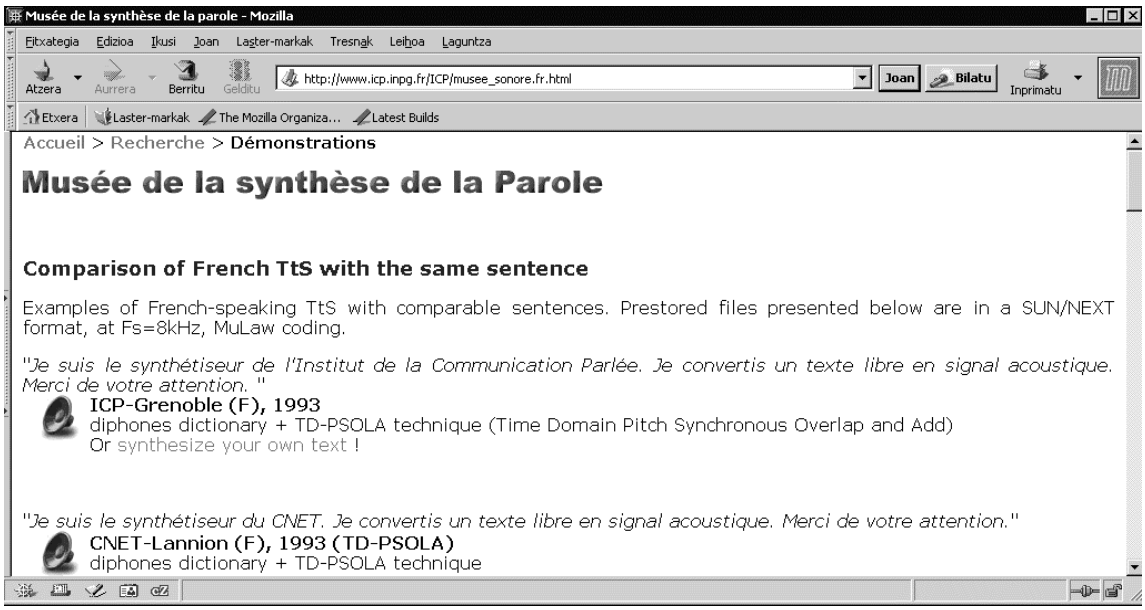
Hiru teknika nagusi daude aipatu dugun helburua erdiesteko: sintesia, analisia eta elkarrizketa. Audio-seinale fisikoaren tratamendua ere garrantzizkoa da, eta beti egiten da aurretratamendu gisa.

Sintesiaren kasuan oso garrantzitsua da adibideak entzutea, hartara jabetu ahal izateko gaur egun noraino iritsi diren eta nola joan diren eboluzionatzen sistemak.

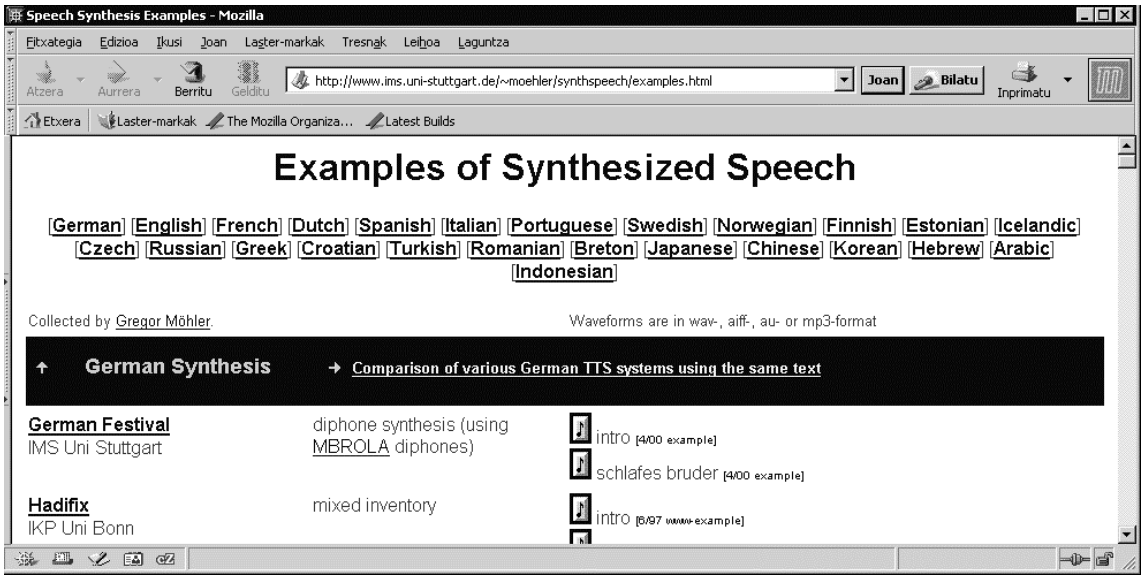
Ondorengo helbideetan, zenbait sintesi- eta analisi-sistema ikus daitezke:

- Linguistic Data Consortium/ COCOSDA Interactive Speech Synthetizer Comparison Site: http://morph.ldc.upenn.edu/ltts/
- Le musée de la synthèse de la parole, Institut de la Communication Parlée, Grenoble (ikus 40. irudia). http://www.icp.inpg.fr/ICP/musee_sonore.fr.html
- EHUko Aholab taldea. Euskara tratatzen duena (ikus 42. irudia). http://bips.bi.ehu.es/ahoweb
- Scansoft, mundu mailan enpresa handiena hizketa lantzeko sistemetan eta hainbat hizkuntza landu dituena. Euskara ere bai. (ikus 44. irudia) http://www.scansoft.com/speechworks/realspeak/demo/default.asp

- Examples of Synthesized Speech on WWW, IMS, University of Stuttgart (ikus 41. irudia). http://www.ims.uni-stuttgart.de/phonetik/gregor/synthspeech/examples.html
- Técnicas y herramientas para el tratamiento automático del habla (ikus 43. irudia) http://liceu.uab.es/~joaquim/language_technology/HLT/tecnol_ling_habla.html

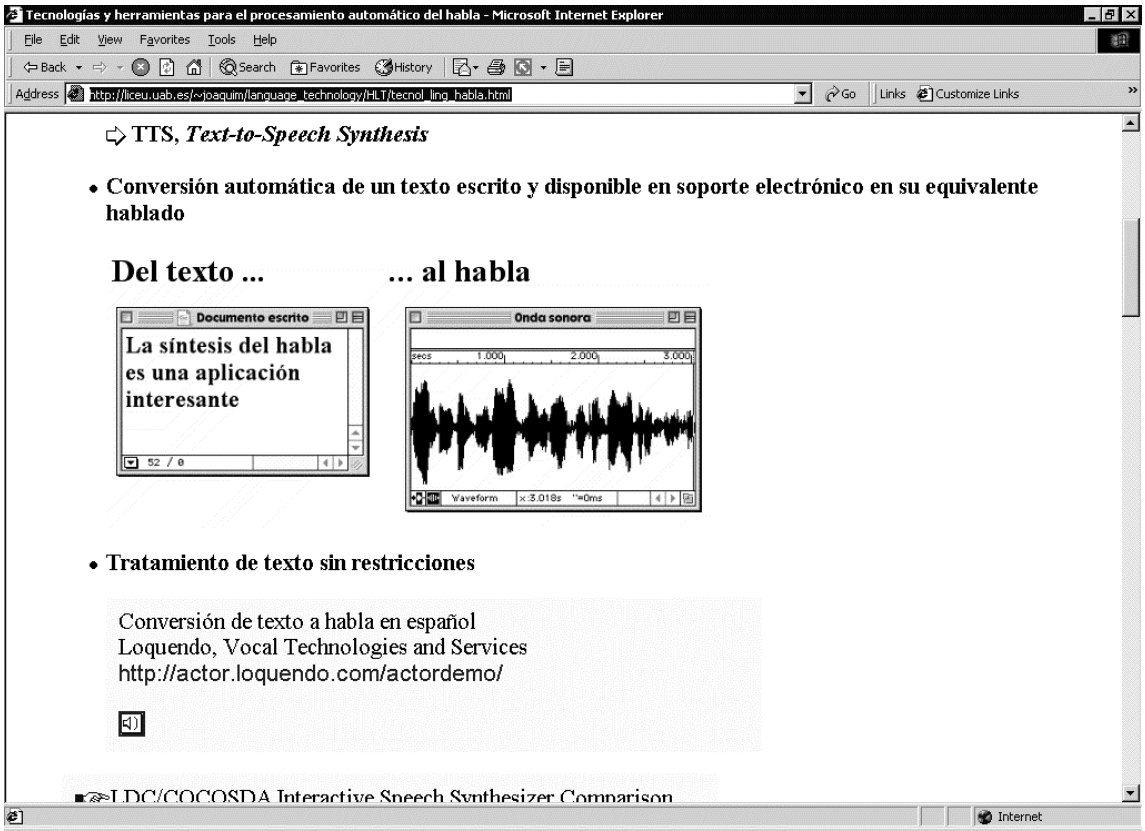


40. irudia. Institut de la Communication Parlée, Grenoble: Le musée de la synthèse de la parole



41. irudia. IMS, University of Stuttgart: Hizketa sintetizatuaren adibideak

42. irudia. AHOLAB Taldea, UPV-EHU



43. irudia. Bartzelonako Unibertsitate Autonomoa: Hizketaren tratamendurako teknikak

Analisia eta elkarrizketaren kasuan, proiektu edota produktu zehatzak ikustea da azalpenik onena zertaz ari garen jakiteko. Horien guztien inguruan, entzunezko, irudizko edota multimediazko adibideak ikusi ahal izateko ere amaraunera jotzea ezinbestekoa dugu.

Zertan dira aipatu hiru teknika nagusi horiek?

- **Hizketaren sintesia.** Ahozko mezuak automatikoki sortzean datza hizketaren sintesia. Ahotsa sistema informatikoak berak sortuko du aurretik zehaztutako datu multzoa edota erregelak baliatuz. Sintesirako hainbat estrategia daude.
- **Hizketaren analisia.** Sintesiaren kontrako teknika dugu. Seinale akustiko bat sistema informatiko batek interpreta dezakeen adierazpide sinboliko bihurtzean datza. Analisi-sistemak honako irizpideen arabera bereizten dira: trata ditzaketen enuntziatuen arabera, onartzen dituzten hiztunen arabera eta prozesa dezaketen hiztegiaren tamainaren arabera. Hitz egiten duenaren ezagupena eta egiaztapena, hizkuntzaren ezagupen automatikoa, eta hizkuntza mintzatuaren ulermena, ahotsaren analisiarekin oso harreman estua duten hiru lan-arlo dira.

  Hitz isolatuen ezagupena erabilgarria da ingurune industrialetan, sistema informatikoen kontrolean edota ofimatikako programetan, etxeko ingurunean, makinak lagunduriko gidatzeko sistemetan eta aire-nabigazioan. Aldiz, hizketa jarraituaren ezagupena edota diktaketa automatikoa teknologia egokia da hizkuntza mintzatua testu bihurtu nahi duten aplikazioetarako.
- **Elkarrizketa-sistemak.** Elkarrizketa-sistemek sistema informatiko batekin harremanetan jartzea ahalbidetzen dute, hartara, informazioa eskuratu edota trukeak egin ahal izateko. Horretarako, analisia eta sorkuntza dute integraturik, eta, horrez gain, elkarrizketa kudeatzeko modulu bat. Bestalde, oso garrantzitsua da hizketa-gaiari buruzko ulermen-ahalmena izatea erabiltzaileen beharrei egoki erantzun ahal izateko. Elkarrizketa-sistemak sailkatzeko honako bereizgarri hauek erabiltzen dira: sistemak erabiltzailearekin duen elkarrekintza mota, sistemak egiten duen lan mota, diseinatua izan den egoeretarako ezaugarri espezifikoen mota, eta elkarrizketa bera garatzen den modua. Informazioa bilatu eta eskuratzea eta transakzioak egitea dira elkarrizketa-sistemen ohiko aplikazioak. Horrekin batera, elkarrizketen itzulpen automatikoa ere aipa litezke. Azken horrek LNPko teknologia konplexuagoak erabiltzea eskatzen du, baina oinarrizko zenbait sistema dira erabilgarri.

44. irudia. Euskarazko testu bat irakurtzen duen Scansoft enpresaren orria

7 Tresnak

Jarraian, Natural Language Registry izeneko erregistrotik (registry.dfki.de) hartuta, egun eskuragarri dauden hainbat tresna aurreratu edo inguruneren zerrenda aurkezten da. Atal honetako tresna guztien helbidea probatu dela esan beharra dago eta gehienak unibertsitateetan garatu dira. Azpiataletan banatu ditugu tresnak espezializazio-arloaren arabera; horretara, atal hauek bereizi ditugu: hizketa, baliabide linguistikoak eta terminologia, morfologia, sintaxia, eta semantika. Zenbait tresna beste atal batean ere sar litezke. Informazio zehatzagoa nahi izanez gero, jo zuzenean Natural Language Registry biltegira.

7.1 Hizketa

AGTK: Annotation Graph Toolkit

AGTK software-osagai multzoa da, seinale linguistikoen bidez edozein portaera linguistikoren berri ematen duten serie kronologikoko datuak (audioa, bideoa) etiketatzeko tresna lagungarriak osatzeko. Barruko datuen egiturak etiketa-grafoetan daude oinarrituta. http://agtk.sourceforge.net

EXMARaLDA

XML lengoaian oinarritutako sistema hizketa ordenagailuz transkribatzeko, JAVA lengoaian idatzitako transkripzio-tresna ere barne (partitura-editorea)
http://weblex.ens-lsh.fr/projects/xitools/index.htm

WaveSurfer

Hizketa bistaratzeko eta transkribatzeko tresna da, oso egokia hizketaren ikerketako eta hezkuntzako hainbat eginbeharretarako. http://www.speech.kth.se/wavesurfer/

7.2 Baliabide linguistikoak eta terminologia

AGTK: Annotation Graph Toolkit (aurreko atalean azaldua)

DIMAP

DIMAP-4 sistemak definizioen azterketa egiten du sare semantikoak sortzeko, hitz-adierak desanbiguatzeko aukera ematen du eta galderak erantzuteko gaitasuna du, *Proximity Parser* erabilita. http://www.clres.com/DIMAP.html

Ellogon

Hizkuntza-ingeniaritzako ingurune eleaniztun, orokor eta plataforma anitzetan dabilena da Ellogon. Hizkuntzalaritza konputazionaleko ikertzaileei nahiz hizkuntza-ingeniaritzako sistemak sortzen eta banatzen dituzten enpresei laguntzeko sortutako ingurunea da. Ellogon-ek baliabide ugari eskaintzen ditu; hala nola, *txt*, *html* nahiz *xml* lengoaietan dauden datuak eta horiei lotutako informazio

linguistikoa prozesatzeko eta bistaratzeko tresnak; baliabide lexikaletarako laguntza (esate baterako, lexikoak sortzea eta txertatzea); corpus etiketatuak sortzeko, datu-baseak atzitzeko eta datu etiketatuak alderatzeko tresnak, edo informazio linguistikoa bektoreetan eraldatzeko tresnak, ikasketa automatikoko zenbait algoritmorekin erabiltzeko. http://www.ellogon.org

Fastr (A tool for automatic indexing)

Terminoak eta horien aldaerak ezagutzeko analizatzaile sintaktikoa da *Fastr*. Sarrera gisa corpusa eta termino-zerrenda bat hartzen ditu *Fastr*-ek, eta horren emaitza edo irteera terminoak eta aldaerak ezagututa dauden corpus indexatua da. *Fastr* tresnaren formalismoa PATR-II-tik gertu dago. http://www.limsi.fr/Individu/jacquemi/FASTR

IMSLex

IMSLex hiztegien datu-baseak informazioa ematen du alemanezko hamar milaka oinarrizko formaren flexioei, hitzen osaerari eta balentziei buruz). IMSLex datu-basetik lexiko espezializatua lor daiteke lengoaia naturala prozesatzeko zenbait aplikaziotarako, hala nola informazioa berreskuratzeko eta erauzteko. Beharrezkoa izanez gero, informazio semantikoa ere erants daiteke *Tübingen GermaNet* hiztegi lexiko-semantikotik.
http://www.ims.uni-stuttgart.de/projekte/IMSLex HPSG

INTEX

Garapen-ingurunea da. Ingurune honi esker, hizkuntzalariek estaldura handiko deskribapenak egin ditzakete, hiztegi elektronikoak eta egoera finituko gramatikak, kasurako. Halaber, INTEXen bidez, deskribapen horiek testu handiei aplika dakizkieke (milioika hitz denbora errealean). Prest daude dagoeneko frantses, ingeles, greko, italiera, portuges eta gaztelaniarako estaldura handiko deskribapenak. http://www.nyu.edu/pages/linguistics/intex

PAGE: A Platfrom for Advanced Grammar Engineering

DFKI GmbH Adimen Artifizialeko Alemaniako Ikerketa Zentroko Hizkuntza-Teknologiaren laborategian garatutako eta mantendutako sistema da. LNPrako motor nagusi aurreratua da, batik bat baliabide gramatikalen eta lexikalen garapena errazteko, motadun ezaugarrien logikan oinarrituta (adibidez, HPSGren edo 'guneak zuzendutako egitura sintagmatikoen gramatiken' markoetarako).

PAGE sistema arrakastatsua izan da gramatikak garatzeko nahiz aplikazioak sortzeko, DFKIn bertan nahiz kanpoan. Egiaztatu da plataforma heldu eta sendoa dela eskala handiko ingeniaritza konputazionalerako. PAGE sistema erabili da estaldura handiko alemaneko, ingeleseko eta japonierako gramatikak egiteko. http://www.dfki.de/pas/f2w.cgi?lts/page-e

ThoughtTreasure

ThoughtTreasure lengoaia naturala prozesatzeko 'sen onaren' ezagutzaren datu-base eta arkitektura da. Hainbat adierazpide erabiltzen ditu, hala nola, logika, automata finituak, saretak eta scriptak

http://www.signiform.com/tt/htm/tt.htm. ThoughtTreasure ezagutzaren datu-baseak osagai hauek ditu: 35.000 hitz eta sintagma, ingelesez; 21.000 hitz eta sintagma, frantsesez; 27.000 kontzeptu eta kontzeptu horiei buruzko sen oneko 51.000 asertzio.

ThoughtTreasure datu-basearen arkitekturak elementu hauek ditu: testu-agentzia bat hitzak, sintagmak eta izendatutako entitateak testuan etiketatzeko; zuhaitz sintaktikoak sortzeko osagai sintaktikoa; azaleko azterketa semantikoak egiteko eta anaforak ebazteko osagai semantikoa; sortzaile bat asertzioak ingelesera eta frantsesera bihurtzeko; plangintza-agentzia bat mundu simulatuan helburuak lortzeko; eta ulermen-agentzia bat testuak sortzeko eta sakoneko ulermenerako.

7.3 Morfologia

AGTK: Annotation Graph Toolkit (aurreko atalean azaldua)

Alvey Tools Grammar Development Environment

Alvey Tools GDE ingurune sendo eta heldua da baterakuntzan oinarritutako gramatika handiak garatzeko. GPSG estiloko estaldura handiko ingeleseko gramatika egiteko erabili zen jatorriz. Analizatzaile sintaktiko eraginkorra eta analisiaren emaitzak ikusteko nabigazio-tresnak ere baditu. Horrez gain, ingeleseko zenbait gramatika eta lexiko zabala ere eskaintzen ditu ingurune horrek. http://www.cl.cam.ac.uk/Research/NL/anlt.html

CUF: Comprehensive Unification Formalism

CUF murriztapenetan oinarritutako lengoaia da, deskribapen linguistikoak zehazteko eta prozesatzeko (sintaxia, fonologia, morfologia eta semantika).
http://www.ims.uni-stuttgart.de/projekte/cuf

*Ellogon (*aurreko atalean azaldua)

FSA Utilities

Espresio erregularrak, egoera finituko automatak eta egoera finituko transduktoreak manipulatzeko utilitate sorta. Honelako manipulazioak egin daitezke: automatak osatzea espresio-erregularrak abiapuntu hartuta, determinizazioa (egoera finituko hartzaileentzat nahiz egoera finituko transduktoreentzat), minimizazioa, konposizioa, osaketa, ebakidura, Kleene itxitura eta abar. Zenbait bistaratze-tresna erabil daitezke egoera finituko automatak arakatzeko. Interpretatzaileak ere eskaintzen dira egoera finituko automatetarako. http://odur.let.rug.nl/~vannoord/Fsa

GATE: General Architecture for Text Engineering

Giza hizkuntza prozesatzeko munduko sistemarik erabilienetako bat da GATE. Tresna egokia da giza hizkuntza prozesatzeko esperimentuak egiten diharduten zientzialarientzat, hizkuntza prozesatzeko osagaidun aplikazioak egiten dituzten enpresentzat eta hizkuntzalaritzari eta hizkuntzalaritza konputazioanalari buruzko ikastaroak egiten dituzten irakasleentzat nahiz ikasleentzat.

GATE tresnak arkitektura, markoa (edo SDK) eta garapen-ingurune grafikoa eskaintzen ditu, eta Sheffield-eko Lengoaia Naturalaren Prozesamendurako taldeak prestatu du azken zortzi urteetan. Hizkuntza prozesatzeko hainbat proiektutarako erabili da sistema hori, batik bat hizkuntza askotan egin den informazio-erauzketarako. Hizkuntza prozesatzeko osagaien bizi-ziklo osoa hartzen du aintzat sistemak, hasi corpusa bildu eta etiketatzetik, sistemaren ebaluazioa egiteraino. GATE sistema EPSRC Engineering and Physical Sciences Research Council-ek eta EBk finantzatzen dute. http://gate.ac.uk

INTEX (aurreko atalean azaldua)

JUMAN

JUMAN sistema japonierako analizatzaile morfologiko malgua da eta erabiltzaileek oso erraz pertsonaliza dezakete. Erabiltzaileek oso erraz birdefini ditzakete japonierako gramatika-sistema morfologikoaren definizioak; esate baterako, kategoria, inflexio-arauak eta morfemak lotzeko arauak.
http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman-e.html

LanguageTool

Iturburu irekiko hizkuntza-zuzentzailea da ingeleserako. Arauetan oinarritutako hizkuntza-zuzentzailea da. Erroreak aurkitzeko balio du eta errore bakoitzerako arau bat definitzen da XML konfigurazioko fitxategietan. Estilo- eta gramatika-erroreak detektatzen dituen 40 arau inguru ditu. Gerta daiteke batzuetan esaldi batek arau bat abiaraztea, nahiz eta esaldia zuzena izan. Oraindik arauak ezartzeko eredua behar bezain zehatza ez delako gertatzen da hori.
http://www.danielnaber.de/languagetool

LT CHUNK

LT CHUNK zatitzaile sintaktiko (*chunker*) edo analizatzaile sintaktiko partziala da. Etiketatzaile batek ematen duen kategoriaren informazioa eta neurri batean testuinguruaren mendeko diren gramatikak erabiltzen ditu, talde sintaktikoen mugak detektatzeko. Aurrez testuan erantsitako informazio guztia utzi eta zatiaren hitzak barne hartzen dituen egitura-elementu bat sortzen du zatitzaileak. Oraingoz izen- eta aditz-talde arrunten mugak ezagutzeko gai da.
http://www.ltg.ed.ac.uk/software/chunk/index.html
http://www.dfki.de/~neumann/morphix/morphix.html

MontyTagger v1.0

Arauetan oinarritutako kategoria-etiketatzailea da Monty Tagger. 1994an Eric Brill-ek trasnformazioan oinarritutako ikasketa-sistema duen kategoria-etiketatzailea egin zuen, eta horretan dago Monty Tagger oinarrituta. Brill-en sistemarekin bateragarriak diren lexikoaren eta arauen fitxategiak erabiltzen ditu. (Brill-en jatorrizko *Penn Treebank*-eko lexikoaren eta arauen fitxategiak ere banatzen dira.) Horrez gain, ingeleserako tokenizatzailea eta performantzia aztertzeko tresnak ere eskaintzen dira. TVersion 2.0 delakoak hibridatu egingo du Brill-en etiketatzea linguistikoki motibatutako teknikekin; horren helburua da ingelesezko kategoria-etiketatzearen doitasuna

handitzea, alderdi hauetan hain zuzen ere: *phrasal verbs* direlakoen eta adierazpide idiomatikoen tratamendua, azaleko analisia, distantzia handiko mendekotasunen tratamendua, aditzen hautatze-lehentasunak eta sen onaren hautatze-lehentasunak.
http://web.media.mit.edu/~hugo/montytagger/

MORPHIX

Alemanerako oso osagai morfologiko azkarra eta sendoa da Morphix. Flexioaz gain, osagaiak ere aztertzen ditu. Halaber, hitz-formak ere sor ditzake lema-sarrera jakin batetik abiatuta, eta informazio morfosintaktiko gehiago ere ematen du.

PAGE (aurreko atalean azaldua)

The Quipu Grok Library:

Java osagaien biblioteka da Grok, lengoaia naturaleko zenbait funtzio betetzeko. Esate baterako, prozesatu aurreko hainbat egikizun egiteko (kategoria etiketatzea, esaldiak detektatzea eta abar), diagrama bidezko azterketak egiteko, kategorien araberako gramatika zabala ingeleserako egiteko (*Penn treebank*-etik induzitutakoa) eta ezagutza adierazteko zenbait osagai sortzeko (oinarrizko korreferentzia, salientziaren jarraipena eta abar). http://grok.sourceforge.net

7.4 Sintaxia

AGFL

AGFL (*Affix Grammars over Finite Lattices*) testuingururik gabeko gramatikak deskribatzeko formalismoa da. Bi mailako gramatikak dira: testuingururik gabeko lehen maila handituz doa, kategorien arteko bat etortzea adierazten duten elementuen bidez. http://www.cs.kun.nl/agfl/

AGTK: Annotation Graph Toolkit (aurreko atalean azaldua)

ALE:Attribute-Logic Enginee

ALE 3.2 Prolog lengoaian idatzitako doako programa da eta elementu hauek hartzen ditu bere baitan: sintagmen egituraren analizatzailea, guneak zuzendutako sortzaile semantikoa eta murriztapen-programazioa, termino gisa motadun ezaugarri-egiturak dituena. Horrek PATR-IIren ezaugarri-egiturak eta Prolog IIren terminoak konbinatzen ditu, mota-herentzia eta mota-egokitasunari buruzko zehaztapenak egin ahal izateko, ezaugarrietarako zein balioetarako. Motei murriztapen arbitrarioak ezar dakizkieke eta egitura-baldintza gehigarriak izan ditzakete. Gramatikek baterakuntza-pausoak programaren deiekin tarteka ditzakete (klausula zehaztuen gramatiketan –DCG– egin daitekeen bezala). Hartara, sistemaren beste osagai batzuekin tarteka daiteke azterketa sintaktikoa. HPSG guneak zuzendutako egitura sintagmatikoen gramatikari begira garatu zen hasieran ALE, baina PATR-II gramatikak, klausula zehaztuen gramatikak (DCG), Prolog, Prolog II eta LOGIN

programak ere exekuta ditzake. Kodetze egokiarekin LFG gramatika lexiko-funtzionalaren zenbait alderdi ere exekuta ditzake. http://www.cs.toronto.edu/~gpenn/ale.html

Alvey Tools Grammar Development Environment (aurreko atalean azalduta)

Annotate

Etiketatze sintaktikorako tresna grafikoa da. Negra corpusa egiteko erabili da. Egitura sintaktikoa ezartzeko balio du eta nodoetan dauden elementuei buruzko informazioa ematen du. Etiketatzeko lana errazten du, interfaze erraz eta erabilerraza prestatu baita. Etiketa egokiak proposatzen dituen kanpoko etiketatzaile/analizatzaile bat dauka interfazeak, metodo probabilistikoetan oinarritua. Etiketatzen diharduen pertsonak proposamena onartzeko edo baztertzeko aukera du. http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html

*CUF (*aurreko atalean azaldua)

Dggraph

PostScripteko biblioteka da Dggraph. Mendekotasun-grafo baten zehaztapenetik (adibidez, zuhaitz sintaktiko batetik) abiatuta, arku-grafo bat eraikitzen du. PostScript fitxategia beste formatu grafiko batzuetara alda daiteke, esate baterako EPS/EPSF, PS, PDF, PNG, GIF eta JPEG formatuetara. Dggraph-en bidez arku gurutzatuak (etenak) egin daitezke. Nodo bakoitzak etiketa kopuru arbitrarioa izan dezake, bata bestearen azpian jarrita, eta bi arku multzo egiteko aukera ematen du: bata nodoen gainekoa, eta nodoen azpikoa bestea. Grafoa lerro edo orri batean sartzen ez bada, behar adina lerrotan edo orritan banatzen du. http://www.id.cbs.dk/~mtk/DGgraph

Ellogon (aurreko atalean azaldua)

ELSPS

Analizatzaileen eraginkortasuna enpirikoki alderatzeko metodologia eskaintzen du ELSPSk. Gramatika, lexiko eta proba-programetarako estekak ere baditu.
http://www.informatics.susx.ac.uk/research/nlp/carroll/elsps.html

EVALB

Parentesiak ezartzeko zehaztasuna ebaluatzen du. Parentesi-ezartzearen doitasuna eta estalduraren berri ematen du, eta etiketatze–zehaztasuna ere neurtzen du. http://nlp.cs.nyu.edu/evalb

FSA Utilities (aurreko atalean azaldua)

*GATE (*aurreko atalean azaldua)

INTEX (aurreko atalean azaldua)

LanguageTool (aurreko atalean azaldua)

LT TTT

Testuak tokenizatzeko eta hainbat mailatan etiketatzeko baliabide malgua eskaintzen du Lt TTT sistemak. LT TTT sistemaren osagai nagusia helburu orokorreko kaskada-transduktorea da. Transduktore horrek era deterministan prozesatzen du sarrera-korrontea eta berriz idazten du, gramatikako fitxategian jasotako arau sortaren arabera. Sistemak bi osagai ditu, entropia handieneko metodoarekin entrenatutakoak biak ere. Lehena, hitzei kategoria sintaktikoko etiketak esleitzen dizkien etiketatzailea da; eta bigarrena, esaldi-mugen desanbiguatzailea, puntua laburdura baten puntua den edo esaldiaren amaierakoa den erabakitzen duena. http://www.ltg.ed.ac.uk/software/ttt

LX-chunker

Portugeseko esaldien mugak <s>...</s> ikurrekin markatzen ditu eta paragrafoen mugak ... ikurrekin. Portugesez idatzitako testuetako elkarrizketa idatzien konbentzio estandarrez arduratzen da. http://lxsuite.di.fc.ul.pt

MontyTagger v1.0 (aurreko atalean azaldua)

PAGE (aurreko atalean azaldua)

The Quipu Grok Library *(aurreko atalean azaldua)*

TIGERSearch

TIGERSearch softwarea sintaktikoki etiketatutako corpusetarako bilaketa-motor espezializatua da (*treebank* direlakoak). Ingurune nagusiekin erabil daiteke: Windows, Linux, Solaris eta Mac OS X. http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch

7.5 Semantika

DIMAP (aurreko atalean azaldua)

GATE: (aurreko atalean azaldua)

The Quipu Grok Library (aurreko atalean azaldua)

ThoughtTreasure (aurreko atalean azaldua)

8 Baliabide linguistikoak

8.1 Sarrera

Lengoaia Naturalaren Prozesamenduaren (LNP) komunitatean 1980ko hamarkadan baliabide lexikal zabal eta aberatsen beharra zabaldu zen. Alderdi teorikotik, teoria linguistiko erabilienak (segur aski Chomsky-rengandik hasita, 1970ean) lexikora lerratzen dira. Aplikazioen ikuspegitik, LNPrako aplikazio errealak garatzeko, ezinbestekoa zen lexiko zabalak edukitzea. Lexikoi konputazionalak dira LNPko aplikazioetako osagai garrantzitsuenetariko batzuk. Lexikoi terminoa erabiltzen da LNPren arloan informazio lexikalaren biltegiei edota hiztegiei erreferentzia egiteko.

Bestalde, esan gabe doa corpusek sekulako garrantzia dutela batetik, hizkuntza alderdi teorikotik aztertzerakoan, eta, bestetik, ikuspegi aplikatutik, informazio lexikala eskuratzerakoan. Testu-biltegiak dira, eta hizkuntzaren prozesamenduari dagokionez, lagin-bankuak. Metodo enpirikoetarako aukera ematen dute.

Laburbilduz, atal honetan gai hauek jorratuko ditugu:

- Sarrera gisa: lexikoari aitortu zaion garrantziak eragindako ekimenak, lexikoaren berrerabilgarritasunak hartu duen indarra eta lexikografia konputazionalak jokatu duen papera.
- Ingeniaritza lexikalaren historiako norabideak.
- Baliabide lexikalak: gainbegiratua.
- Informazio lexikalaren giltzarriak: estandarizazioa, eskurapena eta errepresentazioa.
- Corpusen linguistika.

8.1.1 Baliabide lexikalen beharra Lengoaia Naturalaren Prozesamenduan

1986ko Grosseto-ko mintegia (“Automating the Lexicon”) mugarri garrantzitsua dugu zeren baliabide lexikalekiko kontzientziak esnatzea eta horien garrantziaz jabetzea lortu baitzuen. Mintegiaren bukaeran *Manifiesto* dokumentua osatu zen. Dokumentu horretan baliabide lexikalen (corpus eta lexikoien) garrantzia azpimarratu zen, eta zenbait ekintza gomendatu ziren. Ondorengo urteetan alor honetan Europan garatutako hainbat ekimenen oinarriak ezarri ziren. Batez ere, baliabide lexikal **berrerabilgarriak** lortzeko aktibitate handia piztu zuen.

Egun eremu oso aberatsa, konplexua eta azkar aldatzen dena bilakatu da baliabide lexikalen alorra. Antolakuntza eta teknika mailan garatu egin da. Eta ikerkuntza mailan ere, teknologia berrien, metodologien eta tresnen beharra eskatzen du. Bestalde, ezin dugu ahaztu ezinbesteko oinarriak direla LNPrako eta horren aplikazioen garapenerako, zein hizkuntzen industrien etorkizunerako. Antonio Zampollik horrela zioen First International Conference on Language Resources and Evaluation (LREC) kongresuko komunikazioak jasotzen dituen liburukiaren sarreran (1998:XVI):

> *"The choice of the term "Resources", coined rather recently, was intended to capture the idea that large collections of language data and descriptions play, for development of effective NLP systems and their applications, an essential infrastructural role comparable to the role that basic resources such as highways, railways, electrical networks and energy play for the industrial and economical development of a country".*

Datu linguistikoak baliabide preziatuak dira ezagumenduaren gizartean. Baina, baliabide hitzak normalean beste testuinguru batzuetan izan du bere lekua, baliabide naturalak, ur-baliabideak, baliabide ekonomikoak, etab. Baliabide lexikal kontzeptua linguistika konputazionalari zor zaiola esan genezake. Hona hemen zer dioen Swanepoel-ek Gellerstam-en artikuluan (Gellerstam, 1995:58):

> *"The computer systems and tools that are becoming available both to the researcher, the practical lexicographer and the human user are opening up a myriad of possibilities for the presentation and utilization of masses of lexical information".*

Hala ere, kontua ez da soil-soilik zientzietako adituek lexikografoen eskuetan jartzea tresnak. Datu lexikalak lexikografo adituek bildu eta sistematizatu behar dituzte, zeren eta datu lexikalik gabe linguistika konputazionalak ez du izango zer berrerabili. Eta **lexikoia kondizio *sine qua non* da linguistika konputazionalarentzat**.

Calzolari-k “An Overview of Written Language Resources in Europe: a few Reflections, Facts, an a vision” (Calzolari, 1998) artikuluan azpimarratzen ditu baliabide lexikalek Europan duten egoeraren ikuspegia eta hizkuntzen ingeniaritzaren arloan duten oinarrizko papera. Horrez gain, 1980ko hamarkadaren bigarren aldian eta 1990eko hamarkadaren hasieran lexikoien eta baita ere corpusen inguruan Europan garaturiko hogei proiektutik gorako zerrenda dakar. Autorearen iritziz, proiektu horiek ekarpenak egin dituzte, baina ez dute aurreikusitako estrategiarik edota plangintza garbirik. Arlo honetan diharduten guztien (herrialde desberdinetakoak, proiektu publiko zein pribatuetakoak) elkarlana bultzatu beharra dagoela diosku. Askotan, proiektu berri bati abiada ematean berregiten baitira baliabide lexikalak, eskuragarri dagoena berrerabili gabe. Eta, are okerrago, proiektuetako baliabide lexikalak ahantzita edota erabili gabe uzten dira maiz.

Egoera honi aurre egiteko, 1990eko hamarkadaren lehen erdian Europako Erkidegoko batzorde batek hiru baldintza aipatzen ditu baliabide lexikalen oinarrizko paper hori sendotzeko:

- Baliabide lexikalen eraikuntza zabalkiro onarturiko estandarretan egin beharra.
- Europako Erkidegoko hizkuntza guztietarako baliagarri izango diren baliabide lexikal oinarrizkoen eraikuntza, adosturiko diseinu batez eraikiz.
- Sorturiko baliabide lexikalak komunitateak eskuragarri izan ditzan, distribuziorako politika baten beharra.

Egun Europan arlo honetan dauden proiektu garrantzitsuenek, hain zuzen ere, hiru alderdi horiek lantzea dute helburu batez ere. Proiektu horiek guztiak LE Programme (Language Engineering Programme) barruan kokatzen dira. Hona hemen proiektu horiek zein diren:

- EAGLES (*Expert Advisory Group on Language Engineering Standards*)[18], helburu nagusitzat du estandarren garapena arlo hauetarako:
    - Baliabide lexikal zabaletarako. Adibidez, corpusak, lexikoi konputazionalak, etab.
    - Gomendioak ezagutza lexikal hori erabiltzeko, linguistika konputazionalaren formalismoen bidez, markatze-lengoaien eta zenbait software-tresnaren bidez.
    - Baliabide, tresna eta produktuen ebaluaketarako eta gomendioetarako.
- PAROLE (*Preparatory Action for Linguistic Resources Organisation for Language Engineering*). Proiektu honen helburua da hasierako corpus eta lexikoiak sortzea Europako Erkidegoko hizkuntza guztietarako. Horren ondorik, SIMPLE proiektua aipatzen da PAROLEren jarraipen bezala, eta eginkizun gisa jada existitzen diren geruza morfologiko eta sintaktikoei semantika gehitzea izango du. Bestalde, Europako Erkidegoko hizkuntza guztietarako baliabide lexikalak sortzeko zeregin horretan, EuroWordnet nabarmentzen da. Bere xedea da datu-base eleanitza sortzea hitzen arteko oinarrizko erlazio semantikoak jasoaz.
- TELRI[19], PAROLEren pareko beste proiektu bat. TELRIk (*Trans-European Language Resources Infrastructure*) hizkuntza eta hizkuntzaren teknologia kontuetan puntan dabiltzan zentroen arteko azpiegitura sortu eta industria, ikertoki eta unibertsitateetako LNP komunitateari hizkuntza-baliabide elebakar eta eleanitzak eskaintzea du helburuetako bat. Baliabide horien artean, corpusak, hiztegiak eta lexikoiak euskarri elektronikoan, datu-base lexikalak, eta hizkuntza-datuak sortu, berrerabili, mantendu eta ustiatzeko software-tresnak aipatzen dituzte.
- ELRAk (*European Language Resources Association*)[20], baliabide lexikalak gordetzeko, horien erabilera bultzatzeko eta eraginkorki banatzeko lanak bultzatuko ditu. Honen parekoa Estatu Batuetan LDC (*Linguistic Data Consortium*)[21] dugu.

ELRAk antolatu zuen 1998an LREC (*First International Conference on Language Resources & Evaluation*) kongresua. Eta kongresu hori antolatzen eta bultzatzen parte hartu zuten elkarte (mota askotarikoak) eta errepresentazioa izan zuten herrialde eta hizkuntza anitzek erakusten dute Grosseto-n ernetako baliabide lexikalekiko ardurak hazkunde oparoa izan duela eta etorkizunari begira ere oinarrizko arloa dela.

Arlo honen inguruko informazio interesgarria bezain zabala jasotzen dute ELRAk antolaturiko kongresuetako (Atenasen (2000), Las Palmasen (2002) eta Lisboan (2004)) liburukiek.

8.1.2 Lexikografia konputazionala

Atal honetan lexikografia konputazionalari sarrera labur bat eskainiko diogu.

---

[18] http://www.ilc.pi.cnr.it/EAGLES/home.html

[19] http://www.ids-mannheim.de/telri/telri.html

[20] http://www.icp.grnet.fr/ELRA/home.html

[21] http://www.ldc.upenn.edu/

Lexikografia konputazionalaren alorrean hainbat alderdi bildu ohi dira: ordenagailuz lagunduriko lexikografia (CAL, *Computer Aided Lexicography*), hiztegien edukia automatikoki analizatu eta arakatzeko metodologia eta tresnak, hiztegien azterketa hizkuntzaren egitura semantikoa ikertzeko, eta LNPrako lexiko-sistemen eraikuntza eta horretarako laguntza automatikoak. Horiexek dira, besteak beste, lexikografia konputazionalaren baitan jorratzen diren bideetariko batzuk. LNPren aldetik egindako lexikografia konputazionalaren ikerrarloaren ikuspegi orokorra ematen zaigu *Computational Lexicography for Natural Language Processing* (Boguraev eta Briscoe, 1989) izenekoan. *Longman Dictionary of Contemporary English* hiztegiaren (LDOCE) gainean egindako lanak besterik aurkitzen ez bada ere, deskribatutako teknikak nahiz zenbait lanetan ateratako ondorioak interes orokorrekotzat har genitzake.

Informatikaren aroak aldaketa sakona ekarri du lexikografiaren mundura. Informatika dela medio, astindua ere ezagutu du hiztegigintzak azken urteotan: ederki ahantzia dago paperezko fitxekin lan egiten zeneko sasoia. Gaur, ordenagailua da lexikografiaren lanabes arruntena. Eta, hori horrela, ordenagailuz irakurgarriak diren hiztegiak (MRDak) sortu eta lantzen dira gehienbat. Ia ateratzen diren hiztegi guztiak ateratzen dira euskarri elektronikoren batean (CD-ROMean, batez ere).

Ordenagailuen erabilera lexikografia klasikoan lan hauetara zuzendu da: datuen grabazioa, corpusen lanketarako konkordantzia-programak, testuetako informazio-bilketarako sistemak, hiztegientzako laguntza-inguruneak, etab. Ordenagailuz lagunduriko lexikografiaren arlo honetako lanen artean, Collins argitaletxearen COBUILD hiztegia (Sinclair, 1987) edo 1984an hasitako *New Oxford English Dictionary* (OED) proiektuaren inguruko hainbat lan aipa daitezke (Simpson, 1985; Weiner, 1989).

Bestalde, corpusa oinarri gisa hartzen duten hiztegigintzarako hurbilpenek ez dute tresna konputazional gehiegi izan. Hurbilpen honetako ezaugarri nagusiak honako hauek dira: informazioaren eskurapena corpusetik egingo da; errepresentazio eta modelizazio formala, eta informazio lexikalaren erabilera gizakiari zein LNPko helburuei begira egingo da. (Heid, 1994) lanean jasotzen den legez, COBUILD-ekoak izan dira arlo honetan gehien lan egin dutenak, adibidez HECTOR proiektua garatu zuten. Ildo beretik, eta aipaturiko hutsune hori betetzeko, DELIS proiektua garatu zela aipatzen du Heid-ek. Proiektu horretan corpus-azterketarako eta lexikoi-eraikuntzarako tresnak diseinatu, inplementatu eta integratu ziren ingurune berean.

Corpusa eta hiztegigintza lotzeko tresnen garapenaz aparte, hiztegien mantenimendurako zein sorkuntzarako, datu-baseek ere oso paper garrantzitsua dute hiztegigintzan. Esate baterako, 1992tik 1995era Van Dale Lexicographic Information System (VLIS) proiektua garatu zuen Van Dale argitaletxeak, datu-base eleanitza eraikiz hiztegi elebakarren zein eleanitzen produkzioa eta mantenimendua ahalbidetzeko.

CD-ROMean argitaratutakoek, hiztegi datu-baseak izanik, kontsulta-aukera zabalak eskaintzen dizkiote jendeari. Gai berezituetako terminologi bankuak ere oso tresna baliagarriak dira; horien artean aipa ditzakegu TERMIUM (Kanadako terminologi bankua), UZEIk prestatutako EuskalTerm, etab.

Badira, harantzago joanez, hitza ideiatik abiatuz zein alderantziz bilatzeko aukera ematen dutenak ere. Esate baterako, MEMODATAk (Caen, Frantzia) salgai jarri zuen duela urte batzuk DICOLOGIQUE izeneko hiztegi interaktiboa.

Bestalde, sarreran aipatu bezala, lexikoiak LNPn gero eta garrantzi handiagoa du. Ikertzailea (arlo askotakoak: linguistika, linguistika konputazionala, adimen artifiziala, psikolinguistika, etab.), hizkuntzen industria zein instituzioen interesa lexikora lerratuko dira. Horrela, bada, lexikoaren inguruan hainbat ekintza sustatuko dira: mintegiak, kongresuak, argitalpenak, lan-talde berezituak, etab. (Walker, 1989) lanean zerrenda osoa dugu. Ez da harritzekoa, beraz, lexikoi horien eraikuntza izatea arlo honetako gairik harrotuenetariko bat.

Beste batzuen artean, Europako Erkidegoak martxan jarririko Framework Research Program (1987tik 1991ra garatzen da) aipa dezakegu, zeinak bi helburu nagusi zituen:

- LNPrako ezagutza lexikala errepresentatzeko datu-base lexikalen eraikuntza, beste batzuen artean honako zentroetan ari ziren: Pisako Unibertsitatea (Calzolari eta Zampolli) eta *I.B.M. T.J. Watson Research Center* (Byrd eta beste, 1987).
- Dauden baliabide lexikalak berrerabiltzea, hiztegietan dagoen ezagutza arakatuz eta, batez ere, euskarri magnetikoan daudenez baliatuz, LNPko sistemetako ezagutza lexikala osatzeko. Robert Amsler-ek uste zuen lexikografia konputazionalaren hirugarren aroa hasita zegoela, eta aro berri horretan berrerabilgarritasuna oinarrizko ezaugarria izango zela (Amsler, 1989).

LNPrako lexiko-sistemak eraikitzeari gagozkiola, aurrerago ikusi dugunez LNPko aplikazioen problemarik larriena lexikoi konputazional "handia" eraikitzean datza. Lexikoak ere gero eta garrantzi handiagoa hartuko du, eta hainbat ikertzailek aitortuko dio lexikoari LNPko lanetarako lehentasuna, hiztegiak osagairik behinenak bihurtuko direla. Alderdi teorikotik, hiztegia hizkuntzaren egitura semantikoaren ikerketarako bide den neurrian, zein praktikotik, hiztegiak adimen artifizialaren arazo larriena den jakintzaren eskuratzean izan dezakeen baliagarritasunagatik.

Hori guztia kontuan izanik jabetu gaitezke egun lexikografia konputazionalak hartu duen indarraz, adimen artifizialeko, eta zehazkiago, ezagutzaren errepresentaziorako lengoaien ekarpenez ere laguntza jaso duela, tresneria eta ingurune egokien aldetik batik bat. Lexikoiek testuak interpretatu eta desanbiguatu ahal izateko, adina informazio beharko dute, hizkuntzaren ulermena erdiesteko ahalegin horretan aurrera egin nahi bada.

Lexikoi konputazionalek hiztegi arruntek baino askoz ere informazio linguistiko esplizitu gehiago behar dute, esate baterako, lexikoi konputazional bateko sarrerak gutxienez informazio mota hauetaz hornitu beharko dira: morfologikoa, sintaktikoa (adib. azpikategorizazioa) eta semantikoa (adib. hautapen-murriztapenak), beti ere LNPko sistema batean integratzeko moduan antolaturik. Beraz bi puntu nagusi azpimarratuko genituzke :

- Prozesu automatiko edo erdiautomatikoen beharra, ezagutza lexikalaren eskurapenak lan eskerga eta kostu handikoa baitakar eskuz eginez gero. Horrela, koherentzia eta kontsistentzia ziurragoak izanen dira.
- LNPrako behar diren lexikoiek gero eta errepresentazio-eredu zailagoak eskatzen dituztela.

8.1.3 Datu lexikalak baliabide linguistiko gisa

Baliabide lexikal terminoak (*language resources*, LR) erreferentzia egiten die datu lexikalen multzoei (usu handiak) eta deskribapenei, MRD euskarrian erabiliak izango direnak LNPko sistemak hobetzeko edo ebaluatzeko. Baliabideen artean ditugu: idatzizko zein ahozko corpusak, datu-base lexikalak, gramatikak, terminologia, etab. Egituratzeari erreparatuz gero, bi multzotan bana ditzakegu:

- Baliabide lexikal egituratuak: esate baterako, giza erabiltzaileari zuzenduriko hiztegiak, thesaurusak edota entziklopediak.
- Baliabide lexikal ezegituratuak: corpusa.

Horietaz gain, baliabideen artean sar daitezke baita ere, baliabideen eskurapen, prestaketa, bilketa, kudeaketa eta erabilerarako oinarrizko software-tresnak (Zampolli, 1998).

Baliabide gisa hartzen da, baita ere, introspekzioa, esate baterako, LNPko sistema baterako lexikoia eraikitzen ari den gizakiak hizkuntza eta munduari buruz duen ezagutza.

8.2 Ingeniaritza lexikalaren historiako norabideak

Sintetizatze aldera, LNPrako lexikoen sorkuntzan hiru aro bereiz daitezke:

- Hastapeneko fasea, batez ere, 80ko hamarkadan. MRDetatik lexiko handi eta egokiak eratorriko zirelako ustea.
- Fase goiztiarra, 90eko hamarkadaren hasieran. Estandarren aldarria da nagusi.
- Helduaroa, 95etik aurrerakoa. Gehiegizko optimismoak albo batera utzi eta lan apala eta sendoa egiteko joera gailentzen da.

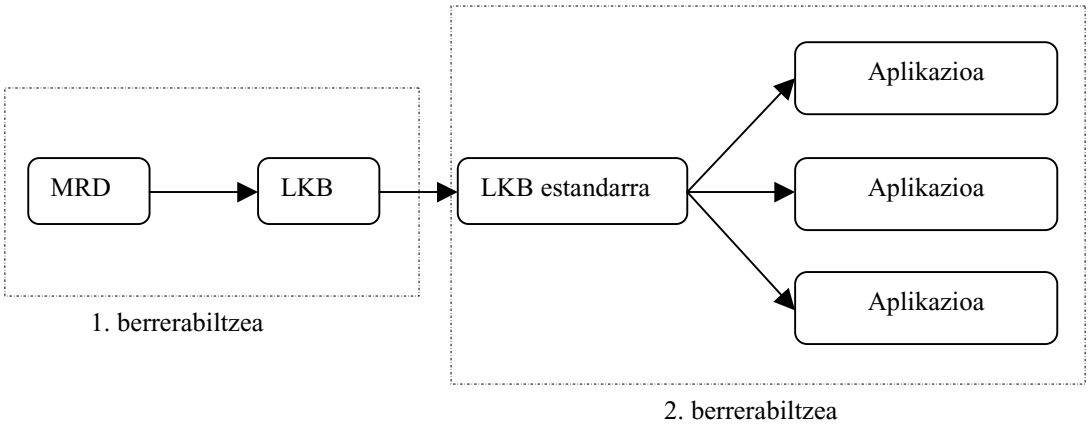
Sintesi gehienek duten ajea du honek ere: sinpleegia da. Baina, halaz guztiz, ingeniaritza lexikalaren bilakaera historikoa ulertzeko marko egokia da.

Bilakaera horrek garbi iradokitzen du zein izan diren ingeniari lexikalen motibazioak. Bi hitzetan labur daitezke motibazio horiek: berrerabilgarritasuna eta estandarizazioa.

Berrerabilgarritasuna, gainera, bi eratara uler daiteke: batetik, hasiera batean giza erabilerarako pentsatuta zeuden baliabide lexikalak LNPrako *berrerabili* ahal dira; bestetik, hiztegien datu-base egokiak diseinatuz gero, hainbat aplikaziotan *berrerabil* daitezke, behin eta berriz informazio bera antolatzen ibili gabe.

Estandarizazioa izan da beste modako hitza hainbat urtetan. Neurri batean, berrerabilgarritasuna lortzeko modua da estandarizazioa, are gehiago, eleaniztasuna ere ustiatu nahi denean. Europan izan du oihartzun handiena joera honek, eta horren erakusgarri dira hainbat proiektu: MULTILEX, GENELEX, EAGLES eta SIMPLE, edota LREC euskarriak. Proiektu horiekin zerikusi estua du TEI (*Text Encoding Initiative*) ekimenak. TEI ez da lexikografiara mugatzen (edozein motatako dokumentuen estandarizazioa lantzen du), baina, arlo honi dagokionez, estandarrak eskaintzen ditu, bai lexikoi konputazionaletarako, bai hiztegi konbentzionaletarako.

Honako eskemak jasotzen du orain artean azaldutakoa:



Hala ere, ingeniaritza lexikorako metodologia horren gainean aldaketa- eta hobekuntza-proposamenak egin dira azken aldian. Besteak beste, honako kezkak plazaratu dira:

- MRDetatik (*Machine Readable Dictionary*) morfologiari eta sintaxiari buruzko informazio lexikala eskuratzen ahal da, baina informazio semantikoak ez du hainbesteko arrakastarik izan.
- MRDak akastunak dira, bai eta erabiltzen diren metodoak ere. Neurri handiko LKB (*Lexical Knowledge Base*) erabilgarriak sortzea zaila da baldintza horietan.
- Denetarako balio duen estandarrik bai ote? Horretan saiatzea baino, ez ote da hobe jatorrizko errepresentazioak mantenduz, estandarrekiko loturak eta egokitzapenak egitea? Izan ere, zaila baita egiturak lotzea, eta gaur-gaurkoz ez baitago semantika lexikalerako estandarrik.

Horiek horrela, aplikazioei orientatutako garapen-metodologia proposatu izan du zenbaitek (Evans & Kilgarriff, 1995). Metodologia horretan ez da planteatzen ezagutza lexikalaren base bat eta bakarra, guztiz berrerabilgarria. Horren ordez, baliabideen integrazio arina proposatzen da, batzuk eta besteak konbinatuz aplikazioek dituzten beharretara egokitzeko modukoak izan daitezen. Estandarizazioa, beraz, azalekoa da, sakonekoa baino gehiago.

8.3 Baliabide lexikalak: gainbegiratua

Ez da samurra baliabide lexikalen mundua zedarritzea. Denetarik da, kasik, *baliabide lexikal* termino generikoaren azpian: euskarri elektronikoko hiztegiak (MRDak), hitz-sareak (WordNet-ak), taxonomiak, terminologia-bankuak, datu-base lexikalak, ezagutza-base lexikalak, ontologiak, LNPrako lexikoiak, corpusak...

Horietako bakoitza zer den zehatz definitzerik ez badago ere, nolabaiteko ezaugarri orokorrak finka daitezke behinik behin. Lau sail nagusitan banatu ditugu:

- Corpusak

- Hiztegiak
- Hiztegi ezagutza-baseak (HEB) eta ezagutza-base lexikalak (EBL)
- Ontologiak

Sailkapen honetako ordena informazioaren elaborazio mailaren arabera egin dugu. Corpusetan hitzei buruzko informazio gordina dago. Hiztegietan, aldiz, lexikografoek kategoria, erabilera-kodeak, definizioa, adibideak, etab. biltzen dituzte. Hitzak ez ezik, hitzen adierak ere azaltzen zaizkigu. HEBetan hiztegietan dagoen informazio inplizitua esplizitu bihurtu eta hitzei buruzko informazio lexikala biltzen da. EBLetan LNPrako sistema batek ulermen eta sormena egiteko hitzei buruz behar duen informazio guztia biltzen dute. Ontologiak munduari buruzko kontzeptualizazioak dira, munduari edo alor konkretu bati buruz jakin beharrekoak (gauza, gertakizun, arrazoinamendu, eta abar, sen ona azken finean) biltzen saiatzen direnak.

8.3.1 Hiztegiak

Hiztegiak, konputagailuz erabiltzeko moduko baliabideak dira. Landugabeak izan ohi dira, testu-prozesadore batez landuak gehienetan. Informazio gordina eduki ohi dute, hori bai, euskarri informatikoan.

Lengoaia Naturalaren Prozesamenduan, 1980ko hamarkadarainoko sistemetan ahaleginaren gehiengoa sintaxi-egituretara eta sintaxitik semantikarako zubietara mugatzen zen. Lexikoa arazorik gabe beteko litzatekeen hitz-zerrenda soil bat besterik izango ez zela uste zen. Garai horretan konturatu ziren LNPrako sistemen hedakuntzarako arazo nagusiena lexikoa urriegia izatea zela, eta lexikoa edukiz betetzea uste baino lan neketsuagoa zela. Garai berdinean, formalismo sintaktiko berri batzuk egitura sintaktikoen pisua lexikoira pasatzen hasi ziren, lexikoaren egitura konplexuago bihurtuz.

Lexiko zabal eta konplexuen eraikuntza eskuz egitea gehiegizko lana izango zela eta, hiztegietan zegoen informazioa ustiatzen ahalegindu ziren. Hiztegi elebakarretan hitzen kategoria, azpikategoria, definizioa, erabilera-adibideak, etab. aurki daitezke. Gainera hitzen esanahiak antolatuta daude, adieren bidez. Berrikiago, hiztegi elebidunetan dagoen informazioa ere ustiatzen hasi da, bai hizkuntza batetik besterako ordainak, baita hizkuntza bateko kolokazio edo eremu semantikoa bezalako informazioa ere.

Hiztegi elebakarren artean, bat izan da tratatua bereziki, *Longman Dictionary of Contemporary English* deritzona (LDOCE, Procter, 1978). Bertako definizioak hiztegi mugatu bat erabiliaz egin dira, ingelesa ikasten ari direnentzat pentsatua. Bestalde, aditzen azpikategorizazioari buruzko informazioa, izenen kode pragmatikoak, arlo semantikoari buruzko kode semantikoak, eta abar jasotzen ditu. Lengoaia naturalaren prozesamenduan aipatzen diren beste hiru hiztegi dira *The Webster's Seventh New Collegiate Dictionary* (Gove, 1969), *Oxford Advanced Learner's Dictionary of Current English* (OALDCE, Hornby, 1974) eta *Collins COBUILD English Language Dictionary* (CED, Sinclair, 1987). Ingelesa ez diren hizkuntzetan hiztegi gutxi tratatu izan dira. Gaztelaniako, adibidez, *Diccionario General Ilustrado de la Lengua Española* (DGILE, Alvar, 1987) eta DREA (Diccionario de la Real Academia Española), CREAn (Corpus de Referencia del Español Actual) oinarritu dira formatu elektronikora pasatu diren batzuk.

Frantseserako *Le Plus Petit Larousse* (LPPL, Larousse, 1980) dago. Euskararen kasuan, *Euskal Hiztegia* (Sarasola, 1997) erabili izan da LNPrako, besteak beste.

Hiztegi hauen erabilera nagusiak, bertatik informazio sintaktikoa erauztea (adibidez, ALVEYko lexikoa horrela eraiki zuten, Boguraev & Briscoe, 1987) eta haiekin HEB edo EBL bat eraikitzea litzateke, hurrengo atalean ikusiko dugun bezala.

Beste hiztegi mota bat *thesaurus*ak dira, sarrerak eduki semantikoaren arabera antolatuta dauzkatenak, aurretik emandako sailkapen bati jarraituz. Lengoaia naturalaren prozesamenduan *Roget's Thesaurus* (Kirkpatrick, 1987) dezente erabili izan da.

Hiztegi elebidunen artean Collins argitaletxeak ingeles-gaztelania, ingeles-frantses, ingeles-italiera, eta abar eskuragarri dauzka formatu elektronikoan. Gaztelania eta ingelesaren artean ere bada *Diccionario Vox/Harrap's Esencial Español-Inglés* (Biblograf, 1992).

Hiztegi moten artean ditugu *Terminologia-bankuak* ere. Hauek termino zientifikoen eta teknikoen gordailuak dira. Terminoen esanahiarekin batera eleaniztasuna lantzen da bereziki. Horren adibide da UZEIk garatutako Euskalterm terminologia-bankua (http://www1.euskadi.net/euskalterm/indice_c.htm)

Erreferentziak

Ondoren, hainbat hiztegiren erreferentziak zerrendatuko ditugu, azalpen labur batzuekin batera:

The Longman Dictionary and Thesaurus:

LDOCE eta LLOCEren MRDak LNPrako lexikoiak sortzeko erabili dira. LDOCE hiztegia era tradizionalean egina dago, baina LNPrako erabilgarriak diren zenbait ezaugarri ditu. Ezaugarri horietarik batzuk, honakoak:

- 45.000 sarrera eta 65.000 adiera ditu.
- 2000 oinarrizko hitz eta horiekin idazten dira definizio guztiak.
- Informazio gramatikal aberatsa.
- Adibideak dauzka.
- Etiketak: erabilera, eremu semantikoa, kode semantikoa (81 daude).

LLOCE hiztegiaren ezaugarriak, berriz:

- LDOCE baino hiztegi txikiagoa da, LDOCEtik eratorria eta printzipio semantikoen arabera antolatua.
- 16.000 sarrera eta 25.000 adiera ditu.
- Hiru mailatan antolatzen da sailkapen semantikoa: Major (14), Group (127), Set (2441).

Adibidea:

<MAJOR: A> Life and living thing

<GROUP: A50-61> Animals/Mammals

<SET: A53> the cat and similar animals: cat, leopard, lion, tiger...

Beste datu-base lexikalekin alderatuz:

- LDOCEk ez du WordNet-ek edo EDRk bezala hierarkia semantiko osorik.
- LLOCEko *clustering*ak ez dira sinonimoenak WordNet-en bezala, erlazionatutako hitzenak baizik
- LLOCEk hiru maila semantiko bakarrik dauzka eta erreferentzia gurutzatuen bidez antzekotasun semantikoa lantzen da.
- LLOCEren klasifikazioa LDOCErena baino landuagoa da. LLOCEk LDOCEren informazio sintaktikoa konbinatzen du thesaurus bateko egitura semantikoarekin.

Oso erabiliak dira LNPn: lexikoi sintaktikoak osatzeko, desanbiguazio semantikorako...

Cambridge International Dictionary of English:

SGMLn kodetua (CIDE+) dago, 80.000 adiera ditu eta 110.000 adibide. http://www.cup.cam.ac.uk/elt/reference/data.htm helbidean topa dezakegu.

LNPrako bertsioak aparteko kodeketa du: domeinua, hautapen-lehentasunak, hitz anitzeko terminoak...

Erabilia izan da dokumentuak laburtzen eta adiera-desanbiguazioan.

GLDB - The Göteborg Lexical Database

Azpian duten modelo linguistikoa lema-lexema da, non lemak forma kanonikoa eta informazio formala (kategoria, flexioa...) errepresentatzen dituen, eta lexemak, berriz, zein sail semantikotan dagoen adierazten du.

GLDBk suediera osorik hartzen du eta bi hiztegi sortu dira bertatik.

Adierazpidean erabiltzen diren erlazio semantikoak: hiperonimia, kohiponimia, hiponimia, sinonimia, oposaketa semantikoa.

Baliabide egokia da, antza, ontologiak eta sare semantikoak sortzeko.

Memodata

Hiztegi-pakete bat bezala defini daiteke: frantsesezko definizioak; frantsesezko, gaztelaniazko, ingelesezko eta alemanezko hiztegi elebidunak, eta frantsesezko ontologia bat ditu.

Frantseserako landua, batez ere. Baliabide guztiak lotuta dauzkate, eta adiera-desanbiguaziorako eta thesaurus moduan erabiltzen da.

Memodata proiektu orokorraren ingurumarian hainbat baliabide landu dira:

- *Dictionnaire integral:* hiztegiaren gunea, bost hizkuntzetarako dago.
- *DICOLOGIQUE:* hiztegiaren bertsio editoriala da.
- *Lexidiom:* hiztegia aldatzeko tresna.
- *Semiographe:* hiztegiaren bertsio konpilatua, aplikazioetan erabil daitekeena

Beste batzuk

Hiztegi arrunten munduan, aurrekoekin batera aipagarriak dira beste hauek ere: *The Webster's Seventh New Collegiate Dictionary (W7), Oxford Advanced Learner's Dictionary of Current English (OALDCE), Collins COBUILD English Language Dictionary (CED).*

Ondorengoak, hiztegi elebidunen erreferentziak dira:

Bilingual Oxford Hachette French dictionary

SGML etiketatua dago, eta etiketa multzo aberatsa dauka.

Collins-Robert English-French dictionary

Hiztegi hau erabiltzen ari da datu-base lexiko-semantiko handi baten eraikuntzan (70.000 kolokazio-bikote).

Ondorengoak, terminologiako hiztegien erreferentziak dira:

Unified Medical Language System

Medikuntzarako lexikoi berezitua. Lau sekziotako antolaketa aberatsa du: Metathesaurus, sare semantikoa, lexikoi berezitua eta informazio-iturburuen mapa.

Eurodicautom

Terminologia-bankua da, web bidez atzigarria. Giza erabilera zabalekoa da. http://europa.eu.int/eurodicautom/Controller helbidean aurki daiteke.

8.3.2 Ontologiak

Ontologiak, munduari buruzko ezagutzaren biltegiak dira. Gizakiok ezagutza hori lexikoaren bidez adierazten dugunez, baliabide lexikalen arloan ere sarri aipatzen dira. Oro har, ezagutzan oinarritutako sistema informatikoek, lengoaia naturala prozesatzen ez badute ere, ontologiaren bat erabiltzen dute.

Ontologiak mundu errealaren kontzeptualizazioak dira, mundu errealari buruzko inferentziak egiteko gaitasuna dutenak. Definizio lauso hau aukeratu dugu, Adimen Artifizialaren arloan definizio zehatzagoek kontrobertsia pizten baitute, eta ontologien ezaugarri bat izango delako guretzat garrantzitsuena: hierarkia darabiltela bizkarrezur. Ontologiak aplikazio askotarako eraiki izan dira (softwarearen berrerabilgarritasuna, medikuntzako sistema adituak, datu-base heterogeneoen integrazioa, lengoaia naturalen sorkuntza, ulermena, itzulpena, eta abar), eta normalean, eremu espezifikoetarako eraiki ohi dira. Hala ere, badira ezagutza orokorragoa biltzen saiatzen direnak ere, adibidez Mikrokosmos, Sensus, CYC, etab.

Autore guztiak daude ados ontologiak oso hetereogeneoak direla esatean, norberaren beharretara neurrira eginak. Hala ere, ontologia guztiek edukitzen dute kontzeptu-zerrenda bat eta kontzeptu horien arteko hierarkia, klase/azpiklase erlazioak egituratuta dagoena. Hori izaten da ontologien ezaugarririk garrantzitsuenetako bat, arestian aipatutako definizio guztietan azaltzen dena.

Ontologien artean ditugu WordNet-ak ere, lehenago ere aipatu ditugunak; literalki itzulita: hitz-sareak. Berez, Princeton-go Unibertsitatean garatutako proiektuak egin zuen ezagun termino hau. Hasiera batean baliabide lexikal jakin hori izendatzeko erabili bazen ere, gaur egun hizkuntza askotarako WordNet-ak garatzen ari dira.

Erreferentziak

Ondoren, hainbat ontologiaren erreferentziak zerrendatzen ditugu:

Cyc, Cycorp

Gizakion sen ona ezagutza-base batean gordetzea du helburu proiektu honek.

100.000 kontzeptu landu dituzte; horietatik 3.000 publikoak dira.

Batez ere webeko informazioaren bilaketan erabili izan da.

Mikrokosmos (Ontos)

Lexikoia eta ontologia bereiz egiten dira. 4.500 kontzeptu biltzen ditu eta, batez beste, 14 erlazio kontzeptuko.

Ezagutzan oinarritutako itzulpen automatikoa erabiltzen da.

Sensus

Makroantolaketa lexiko-semantikoa da. Penman Upper Model, Ontos, LDOCE eta WordNet integratzen ditu.

8.3.3 Ezagutza-base lexikalak eta hiztegi ezagutza-baseak

Ezagutza-base lexikalak (EBL), ezagutzari buruzko informazioa gordetzen duten gordailu egituratuak dira. Ezagutza hau hiztegietatik erauzitakoa denean, hiztegi ezagutza-base (HEB) termino zehatzagoa erabili ohi da. Ezagutza-base edo hiztegi ezagutza-base hauetan, MRDetan ez bezala, entitateak eta beraien arteko erlazioak agerikoak dira, eta normalean semantika lexikala da errepresentatzen dena. Arrazonatzeko eta inferitzeko gaitasuna ere lantzen da.

Ezagutza mota gehienbat gramatikala (kategoria, azpikategoria, morfotaktika...) denean, *datu-base lexikal* (DBL) terminoa erabiltzen da.

Bestetik, LNPrako *lexikoiak* ditugu. Lexikoi terminoak aplikazio batekiko lotura adierazten du. Informazio lexikalaren biltegi hauetan unitate bakoitzari ezaugarri morfologiko, sintaktiko eta semantikoak esleitzen zaizkio; hots, orotariko informazioa maneiatzen dute. Lexikoietan erabiltzen diren errepresentazio-formalismoak sarri teoria jakinetan oinarritzen dira. Ezaugarri-egituren bidezko adierazpidea usu erabiltzen da, eta sistema aurreratuenetan hierarkiak eta herentzia-mekanismoak ere ustiatzen dira.

Horiek guztiak barneratzen ditugu 8.3.3 atal honetan.

Lengoaia naturalen prozesamendu sintaktiko eta semantikoa egin ahal izateko, lexikoiak hitz-zerrenda izatetik EBL izatera pasatu dira, hitz eta adierei buruzko informazioa dutenak. EBL baten hizkuntza ulertu ahal izateko, ordenagailuak hitzei buruz jakin beharreko guztia egon beharko litzateke (Yokoi, 1995). EBLen ezaugarri garrantzitsuena herentzia izaten da, adierak klase/azpiklase hierarkien inguruan antolatzen dira eta (Copestake, 1990). EBLak eskuz eraiki daitezke, adibidez WordNet (Miller *et al.,* 1993b) eta EDR (EDR, 1993), baina askotan hiztegietatik erauzten dira (Copestake, 1990; Bruce *et al.,* 1992).

LNPren beste ikuspuntu batetik, HEBek hiztegietatik erauzitako informazioa jasotzen dute (Artola, 1993). Erauzitako informazioaren artean, hemen ere, adieren hierarkiak dira aipagarriak. HEB batetik EBL bat erator daiteke, hiztegitik zuzenean EBL eraiki daitekeen bezala. HEB baten enfasia hiztegiko informazioan da, inplizitu egon eta esplizitu bihurtu dena, giza erabiltzaileak edo programa batek erabiltzeko moduan. EBL baten enfasia, ordea, LNP aplikazioetarako baliagarria izatea da.

EBL eta HEBak eraikitzeko, hiztegietatik erauzi izan den informazio semantikoa definizioen azterketatik etorri ohi da batez ere, adieren hierarkia eratuz, eta hitzen (edo adieren) arteko bestelako erlazio lexikal-semantikoak finkatuz. Lehenbizi, definizioen analisi sintaktikoa egin behar da, eta ondoren, analisiaren emaitzatik erlazio lexikal-semantikoen erauzketa. Erlazio horietan azaltzen diren hitzen desanbiguazioa ere egin behar da, adieren arteko erlazioak eduki ahal izateko. Horri buruzko zehaztasun gehiago ikusiko ditugu.

Erreferentziak

Ondoren, hainbat ezagutza-base lexikalen erreferentziak zerrendatuko ditugu:

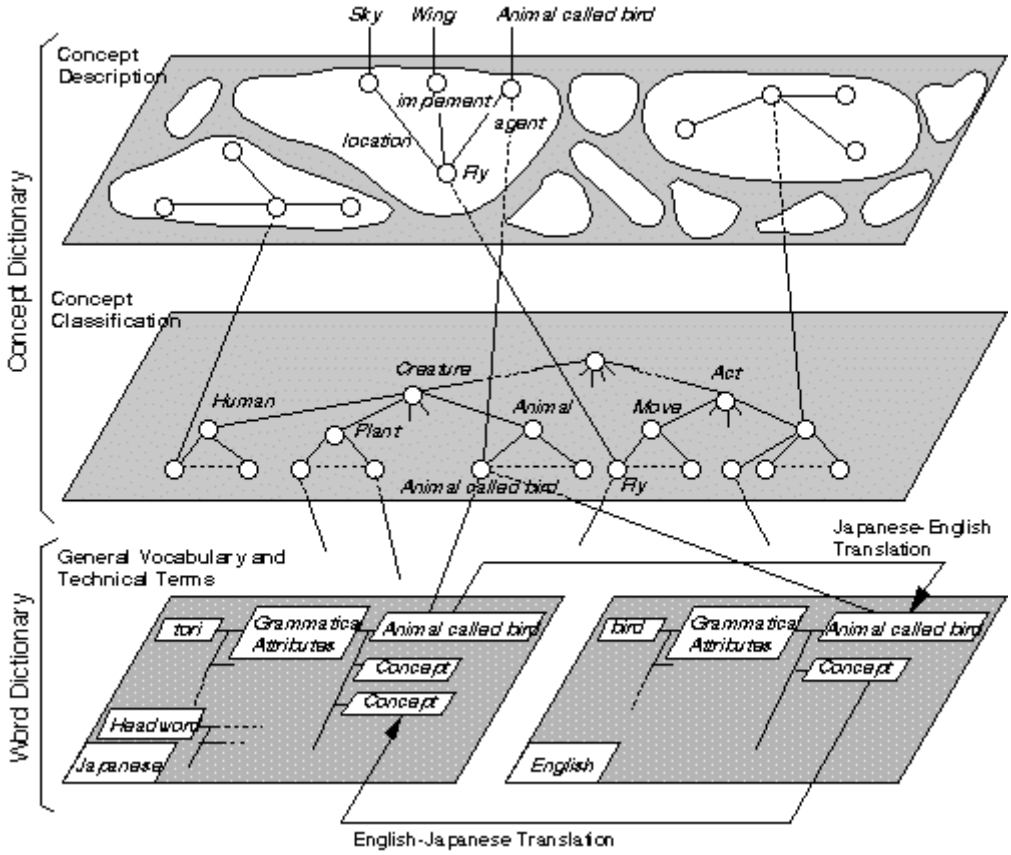
EDR

EDR ingeleserako eta japonierarako hiztegi elektronikoa da, bost hiztegi mota desberdinez osatua dagoena:

- Hitz-hiztegiak (ingelesa eta japoniera): hitz batek errepresentatzen duen kontzeptua azaltzen du.
- Hiztegi elebidunak: japonierazko eta ingelesezko hitzen arteko korrespondentzia deskribatzen dute.
- Kontzeptu-hiztegia: hitz-hiztegian dauden kontzeptuak ulertzeko beharrezkoa da. EDRn kontzeptu bat ulertzeko erabiltzen den errepresentazioa kontzeptuen arteko erlazioetan oinarritzen da.
- Konkurrentzien hiztegia: hitzak adierazteko moduarekin zerikusia duten deskribapenak ematen ditu.
- EDR korpusa.

Beste hiztegietako erreferentziak kontzeptu-hiztegian sailkatu eta erlazionatzen dira.

Izugarrizko baliabidea da, oso aberatsa. WordNet-en antzeko antolakuntza du, adierak kontzeptu-hierarkia batean antolatuta dituelako. Horrez gain, erlazio semantikoen karakterizazio aberatsekoa da.



44. irudia. EDR hiztegi ezagutza-basearen egitura

Itzulpen automatikoan, lexikoaren eskurapenean, eta adiera-desanbiguazioan erabili izan da, besteak beste.

Ondoren, hainbat lexikoi konputazionalen erreferentziak zerrendatuko ditugu:

Corelex

Pustejovsky-ren teoria inplementatzeko ahalegina ("polisemia semantikoa"); hiru pausoko eraikitze-metodologia, non WordNet den abiapuntu; izenak bakarrik landu dira.

Datr

Ezagutza lexikala adierazteko eredu orokorra da. Ekuaziotan oinarrituta dago, eta adierazpen-ahalmen handikoa da. Baterakuntzan oinarritutako gramatikentzat egokia.

8.3.4 Corpusak

Corpusak, testu-biltegiak dira, eta metodo enpirikoetarako aukera ematen dute. Linguistikaren barruan aspalditik izan dira linguistika enpirikoa aldarrikatu dutenak. Horientzat, linguistika ahozko edo idatzizko hizkuntzaren azterketa enpirikoan oinarritu beharko litzateke (McEnery & Wilson, 1996). Idatzizko hizkuntzaren kasuan, azterketaren subjektua idatzizko corpus batek osatzen du. Testu multzo bat corpus izateko, hiru baldintza jartzen dizkiote McEnery eta Wilson-ek: lagin errepresentatiboetan oinarritua

egotea, tamainaz finitua izatea eta makinek tratatzeko modukoa izatea. Corpusek, gainera, errepresentatzen duten lengoaia-aldaeraren erreferentzia estandarra izateko bokazioa eduki beharko lukete. Corpusetan oinarritutako linguistikak 50eko hamarkadan kritika zorrotzak jaso zituenez, jarduera asko murriztu zen. 80ko hamarkadatik aurrera, ordea, onarpen zabala jaso izan du. Zalantzarik gabe, ordenagailuen ahalmena eta makinaz tratatu daitezkeen testuen kopurua etengabe hazten joatea, besteak beste, daude linguistika enpirikoaren berragerpenaren atzean. Gaur egun, linguistikaren alor guztietara zabaldu du bere eragina, ezagutza-baseen aberasketara eta hitzen eta kontzeptuen arteko erlazio-izaeren ikerkuntzara ere ingeleserako erreferentzia-corpus ugari sortu izan dira. Estatubatuarrak izan ziren aitzindari, Brown deritzon corpusarekin (Francis & Kucera, 1967). Britainia Handiko ingelesarentzat ondoren etorri zen London-LUND corpusa (Svartvik, 1990), eta orduz gero etengabe ari dira corpusak berritu, sortu eta aberasten. Corpusean, berez, hitzak besterik ez dago, testu gordinak, baina corpusen erabilera asko zabaltzen da informazio linguistikoa gehitzen badiegu. Informazio hori hitzen kategoria izan daiteke, edo esaldien egitura sintaktikoa (adibidez, Penn Treebank delakoa edo Birmingham-eko *Bank of English* corpusa[22], Murriztapen Gramatiken bidez (Karlsson *et al.,* 1995) kategoria eta egitura sintaktikoz etiketatu dena), edo informazio semantikoa (aurrerago aipatuko dugun *SemCor*[23], hitzen adierez etiketatu den Brown corpusaren azpimultzoa, Miller *et al.,* 1993a). Euskararako, UZEIk sortu du, Euskaltzaindiarentzat, *XX. mendeko euskararen corpus estatistikoa* (www.euskaracorpusa.net), XX. mendeko testuen laginez osatutako 4.650.000 hitzeko corpus estatistikoki lematizatua (Urkia & Sagarna, 1990; Urkia, 2002). IXA taldea euskara estandarra biltzen duen hizkuntza-corpus zabalago bat biltzen ari da.

Kontua da baliabide lexikalen inguruko egitasmo ugari garatu direla azken bi hamarkadetan. Proiektu horietatik batzuk ekarri ditugu hona, ez exhaustibo izateko asmoarekin, baina bai egiten denaren ikuspegi zabal samarra (nahiz azalekoa izan) eskaini nahian.

Egitasmo horiek, zenbaitetan, proiektu estrategikoak izan dira. Horrelakoen artean aipa daitezke, esate baterako, Estatu Batuetan garatutako CYC egitasmoa eta Japongo *EDR Electronic Dictionary*a.

Oinarri konputazional gisa duten interesagatik, baliabide lexikal batzuk aipatuko ditugu segidan.

Erreferentziak

Ondoren, hainbat corpusen erreferentziak zerrendatuko ditugu:

Corpus gordinak

Brown Corpusa: 1 MHitz.

Web bera hartzen ahal da corpus gordin (erraldoi) gisa.

Corpus etiketatuak

Penn Treebank (4,5 Mhitz), *Bank of English (320 Mhitz)*: kategoria eta egitura sintaktikoz etiketatuak.

---

[22] http://titania.cobuild.collins.co.uk/boe_info.html

SemCor: Hitzen adierez etiketatutako Brown-en azpimultzoa.

Euskarazko corpusak

XX. mendeko euskararen corpusa (http://www.euskaracorpusa.net/XXmendea).

Ereduzko prosa gaur : (http://www.ehu.es/euskara-orria/euskara/ereduzkoa/).

EPEC: IXA taldearen corpus orokorra (Aduriz *et al.,* 2003).

Orotariko Hiztegirako erabiltzen ari den corpusa.

Legebiduna: Euskal erakundeetako euskara-gaztelania testu elebidun paraleloak *SGML*/*TEI*-P3 giden arabera etiketatuta (http://www.serv-inf.deusto.es/abaitua/konzeptu/lege2dun.htm

8.4 Informazio lexikalaren giltzarriak: estandarizazioa, eskurapena eta errepresentazioa

8.4.1 LNPko baliabide lexikalak estandarizatzeko premia

Datu linguistikoen berrerabilgarritasuna ziurtatzeko, edozein testuk, konputagailuen laguntzaz aztertuko bada, konputagailuak irakurri ahal izango duen moduan kodeturik egon behar du. Kodeketa estandar baten ezean, eta kodeketa diogunean software eta hardware bateragarriaz ere ari gara, ezinbestean, testuak ikergai dituzten hainbat arlotako ikertzaileek testu horien tratamendua erraztearren hamaika era asmatu eta erabili izan dituzte. Nork berea —eta bere modura— egin duela, ordea, azkenean batek egindakoaz beste bat baliatu nahi izan denean, aurretik egindako lan guztia ez da suertatu izan nahi bezain lagungarria; maiz, erabat erabilezina ere bai. Egoera tamalgarri horren aurrean, saio bat baino gehiago jo izan da azken hogei urte hauetan testuen kodeketarako estandar baten bila, fruituak heldu ez badira ere.

Bestalde, estandarizazioaren beharra ez da unibertsalki konpartitua, horren lekuko, TEI salbu, gainerako estandarizaziorako ekimenak Europan kokatzen direla. Estatu Batuetako linguistika konputazionaleko ikertzaileek uko egin diote estandarren ezarpenari, eta beraien arabera estandarrak praktikatik sortuko dira (behetik gora, *bottom-up*). TEI ekimenekoek sorturiko gidalerroek (maiz TEI P3[24] modura ezagutuak) esfortzu aitzindaria errepresentatzen dute ordura arte ekimen isolatu eta noizbehinkakoak izan ziren arloan. Eta oinarrizko lana dugu formatu elektronikoan dauden testuen kodeketarako etorkizunean. Testu-datuen errepresentazio-arazoak formatu elektronikoan planteatzen dituzte.

1980ko hamarkadako bukaera aldean, oso testu-bilduma egoki gutxi zeuden linguistika konputazionaleko ikerkuntzarako, bereziki ingelesa ez den hizkuntzetarako. Beraz, testu-bilduma erraldoiak (elebakar zein eleaniztun) biltzeko eta zabaltzeko hainbat ekimen sortu ziren, horien artean: *ACL Data Collection*

---

[23] http://www.cogsci.princeton.edu/~wn/

[24] Guidelines for Electronic Text Encoding and Interchange, TEI P3. Amarauneko helbide honetan aurki daiteke gidalerro hauei buruzko informazio zabalagoa: http://www-tei.uic.edu/orgs/tei/p3/elect.html.

Initiative (ACL/DCI)[25], *European Corpus Inititative* (ECI), Estatu Batuetako *Linguistic Data Consortium* (LDC), *Corpus Encoding Standars* (CES), *MULTEXT*[26] Europan, etab. Ekimen horiek guztiak hasiera baino ez ziren, behar zen ahaleginerako, eta baliabide testual handiak egoki eratzeko oraindik lan ikaragarria zegoen egiteko.

Horrez gain, testu-bildumetako berrerabilgarritasunaren eskaerari erantzuteko testu-datuen kodeketarako estandarra garatu beharra zegoen. Datuak formatu *ad hoc*etan zabaltzen jarraitzea errealitateari uko egitea zen, kontuan izanik erabilera partikularretan zer-nolako ahaleginak eta baliabideak behar ziren datuak txukuntzeko eta berrantolatzeko, kasurik hoberenetan kostu handikoak eta askotan ezinezkoak.

Existitzen ziren eta baliagarri izan zitezkeen datu-bilduma gehienak inprimatzeko helburuarekin zeuden formateatuak, eta hori dela eta, kodeketan esplizituki errepresentatutako informazioa gehiago dago lotuta testuaren formatu fisikoari berorren egitura logikoari baino (interesgarriagoa dena LNPko aplikazioetarako); eta gainera, maiz, bien arteko harremana gauzatzea oso zaila edo ezinezkoa izango da lan ikaragarria ez baldin bada egiten.

Horrezaz gain, gero eta datu-bilduma gehiago daudenez eskuragarri, eta datu-bilduma handien erabilera ezinbestekoa denez LNPko ikerkuntzan, orduan, software orokorra eta publikoa testuen lanketarako garatzen ari da berrerabilgarri izateko, eta horretarako kodeketa-formatu estandarra behar da.

Hona hemen Nancy Ide-ren arabera LNPko ikerkuntzarako testuak errepresentatzeko kodeketa-formatu estandar batek bete behar dituen ezaugarriak:

- Gai izan behar da LNPko ikertzaileen komunitatearentzat interesgarri izan daitezkeen testu motak eta hizkuntza anitzetan aurki daitezkeen informazio motak errepresentatzeko, prosan, dokumentu teknikoetan, egunkarietan, poesian, antzerkian, eskutitzetan, hiztegietan, lexikoietan, eta abarretan.
- Informazio maila diferenteak errepresentatu ahal izango ditu, ez bakarrik ezaugarri fisikoak eta egitura logikoa (baita fenomeno konplexuagoak ere, hala nola, testu barnean edo testuen arteko erreferentziak, elementu paraleloen alineamendua, etab.), baita ere, datuei gehi dakizkiekeen anotazio interpretatibo edo analitikoak (esate baterako, kategoria gramatikalen anotazioa, egitura sintaktikoa, etab.).
- Aplikazioarekiko independente, hots, malgutasuna eta orokortasuna izan behar ditu, aldi berean testu bereko informazio motak esplizituki kodetzeko eta prozesamendu mota desberdinetara egokitzeko.

Horrelako kodeketa-sistema malgu eta moldagarria garatzea lan intelektuala da nagusiki, zeinak eskatzen duen testu mota desberdinak jasotzeko modelo konplexuen garapena, testu-modelo orokor bat, eta hori gorpuzteko kodeketa-eskemaren arkitektura.

---

[25] http://www.cs.columbia.edu/~radev/u/db/acl

[26] http://www.lpl.univ-aix.fr/projects/multext

8.4.1.1 TEI: testu-kodeketarako ekimena

Testu-kodeketarako ekimena (*Text Encoding Initiative*, TEI) delakoak testu elektronikoak kodetzeko eta trukatzeko bere gidalerroak —TEI P3 — 1994ko maiatzean eman zituen argitara. Sei urtetan zehar mundu zabaleko hamaika ikertzaile eta ikertalderen lanaren fruitu diren mila eta hirurehun orrialdeko gidalerroen helburua ondokoa bezain sinple eta, aldi berean, handinahikoa da: zenbait ezaugarritako testu mota zabal bat era kontsistente eta hobezin batean kodetzeko bideak eskaintzea.

Gidalerroen garapenean, TEIk identifikatu zituen askotariko ikertzaileek zer-nolako kodetze-premiak zituzten informazioaren trukeari zegokionez, horretan oinarritu zituen orokortu nahi zuen eskema batek bete beharreko kodeketa-printzipioak, eta identifikatu zituen zein ziren kodetze-arauak behar zituzten testu klase eta -ezaugarriak. TEIk eskaintzen dituenetan arakatzen hasita, hona hemen batzuk:

- SGML (*Standard Generalized Markup Language*) markatze-lengoaia egokitzat jotzea gidalerroen garapenerako oinarri gisa.
- SGML erabiltzeko gomendioak —zenbait murriztapen—, bere orokortasuna eta malgutasunari eutsiz aldi berean.
- Testu-datuak kodetzerakoan beharrezko diren kategoria eta ezaugarrien identifikazioa eta analisia, maila askotan.
- Testu-egitura definizio orokorren multzo malgu eta hedagarria.
- Testu elektronikoak dokumentatzeko metodo bat, biblioteketan erabiltzen den katalogatze-arauekin bateragarria.
- Kodetze-arauak testu mota eta ezaugarri desberdinetarako: karaktere multzoak, hizkuntza-corpusak, linguistika orokorra, hiztegiak, datu terminologikoak, ahozko testuak, hipermedia, literatur prosa, olerkia, antzerkia, iturburu historikoak, eta testu-kritikarako aparatua.

Hasiera-hasieratik TEI eskema diseinatu zen hardwarea, software eta aplikazioetatik independente izateko helburuarekin. Aplikazioetatik bereizi-nahi horrek izugarrizko garrantzia du, gure ustez, eta testu baten hainbat ikuspegi kodetu ahal izateko aukera ematen digu. Izan ere, testu bat ikus baitaiteke objektu fisikoen bilduma bezala (liburukiak edo paper-orri solteak), edo objektu tipografikoen segida bezala (karaktere-sekuentzia, letra-molde eta marjina-eskemen arabera antolatuta), edo objektu linguistikoen sekuentzia bezala (grafema edo fonemak, morfemak, unitate lexikalak, sintagmak...), edo objektu formalez osatutako egitura bezala (ahapaldiak, lerroak, kapituluak, atalak...), eta abar eta abar. TEI gidalerroek helburu orokorreko kodetze-eskema bat definitzen dute, ikuspegi horiek guztiak era askotan kodetzeko aukera ematen duena, eta, nahi izanez gero, aldi berean gainera.

Gidalerroek ikerketarako beharrezko diren testu-ezaugarriak errepresentatzeko aukera asko eta asko ematen dituzte, eta, oro har, aitortzen zaie egun diren premia gehienetarako baliagarritasuna. Bestalde, gidalerroen diseinuak berak bere hedagarritasuna bermatzen du, eta, horretara, estandarrean definituriko elementuekin-eta aski ez duen erabiltzaileak aukera guztiak ditu dokumentu motak (zeinak DTD, *Document Type Definition* delako fitxategian definitzen diren) bere premietara egokitzeko (gidalerroetan, gainera, adierazten zaio nola egin gauzak nahasten ibili gabe). Software berezirik gabe erabili ahal izateaz den bezainbatean, berriz, esan beharra dago posible dela editore estandarrez-eta baliatuz lan egitea, baina

aitortu behar da askoz ere emaitza hobeak eta eraginkortasun handiagoa lortuko dela, inondik ere, editore bereziak eta, oro har, software berezitua erabiliz gero.

Bestalde, diseinatutako eskemak erantzun bat ematen die jada kodetze-proiektu gehienen oinarrizko premiei; hala ere, gidalerroak zabaldu egin behar dira eta, batik bat, erabiltzaileak behar-beharrezko diren laguntza-tresnez hornitu behar dira. Izan ere, estandar bat "saltzeko" modurik hoberena estandar horren erabilera erraztuko duten software-tresna eta baliabideak garatzea baita.

Erabiltzaile askoren premiak betetzera datoz TEI P3 gidalerro hauek: zientzia eta giza arloko ikertzaileak, argitaratzaileak, bibliotekariak, eta, oro har, dokumentuen bilaketa eta biltegitzearekin zerikusia duten guztiak. Erantzun bat ematen dio, orobat, "hizkuntzaren teknologiaren" arloko jendeari, orotariko testu-corpus eta lexikoen pilatzeari emanak baitaude azken aldi honetan, hizkuntzaren ulerkuntza, sorkuntza eta itzulpenari dagokion ikerkuntzan sartuak.

8.4.1.2 SGML: testuak markatzeko lengoaia estandar eta orokorra

Testuak markatzeko lengoaia estandar eta orokorra, hots, *Standard Generalized Markup Language* (SGML) marka multzo bat baino areago marka multzoak espezifikatzeko metalengoaia bat da. Eta metalengoaia horretaz baliatuz diseinatu dira TEI gidalerroak.

SGML lengoaian, testu barnean markatze-kodeak txertatzen dira eredu bati jarraituz eta modu deskriptiboan egituratzen da. Programazio-lengoaien edo aplikazioen mende sortzen diren egitura berezituak saihesten dira.

Markatze-sistema deskriptiboa da, eta honelako markatze-sistemek markatze-kodeak erabiltzen dituzte dokumentuaren zatiak izendatzeko. Horrela, bada, <esaldia> edo </esaldia> bezalako kodeak dokumentu batean txertatzen ditugunean, testu horren zati baten hasiera edo bukaera adieraziko genuke.

SGMLk dokumentu motaren nozioa ere ezartzen du, hau da, dokumentu bakoitza mota batekoa izango da eta dokumentu mota batek dokumentu multzo bat definituko du. Dokumentu mota hau DTD (*Document Type Definition*) delako fitxategian definitzen da eta dokumentu guztiek DTD bati esleituta egon behar dute.

DTDan dokumentuaren mota formalki definitzen da, hots, bere osagaiak eta egitura esplizituki adierazten dira. Adibidez, txosten bat definitzerakoan, hasieran egilearen izena etorriko dela esan beharko dugu, ondoren laburpen bat eta, azkenik, txostenaren muina daukaten hainbat paragrafo. Esaten da SGML lengoaia metalengoaia dela, DTDen bidez (edo, SGML terminologia erabilita, "aplikazioen" bidez) azpilengoaiak definitzen baitira. Azpilengoaia horien adibide behinena Interneten-eta hain erabilia den HTML (*Hyper Text Markup Language)* dugu.

Adibide gisa, demagun poemak gordetzeko barne-egitura definitzen dugula. *Antologia* bat hainbat *poema*z osaturik egongo da. *Poemek*, *izenburu* bana eta hainbat *ahapaldi* izango dituzte. *Estrofak* hainbat *lerro*z osaturik egongo dira.

SGML lengoaian kodetuta, egitura horrekin bat datorren testu batek horrelako itxura izango luke:

```
<antologia>
   <poema>
      <izenburua>Potaren galdatzia</izenburua>
         <ahapaldia>
            <lerroa>Andria, Ieinkoak drugatzula; orai berdi girade</lerroa>
            <lerroa>ni errege balin baninz, erregina zinate;</lerroa>
            <lerroa>pot bat, othoi, egidazu; etzaitzula herabe;</lerroa>
            <lerroa>nik zugatik dudan penek hura merexi dute.</lerroa>
         </ahapaldia>
         <ahapaldia>
            <lerroa>Eia horrat, apart´ adi; nor uste duk nizala?</lerroa>
            <lerroa>Horlako bat eztuk uste nik ikusi dudala;</lerroa>
            <lerroa>horrelako hitz gaixtorik niri eztarradala;</lerroa>
            <lerroa>berzer erran albaitzita; enuk uste duiana.</lerroa>
         </ahapaldia>
   </poema>
   <poema>
      <izenburua>...
      ...
   </poema>
      ...
</antologia>
```

Gaur egun, XML (*Extensible Markup Language*) lengoaia erabiltzeko joera hedatzen ari da. Azken batean SGMLtik eratorritako azpimultzo bat da.

8.4.2 Informazio lexikalaren eskurapena

Lexikoi konputazionalak hutsetik edota eskuz eraikitzea zailtasun handikoa izanik, datu lexikalen eskurapenaren arazoari begira "berrerabilgarritasuna" hitz gakoa da. Baliabide lexikalei buruzko eztabaidek eta honen inguruan garatzen ari diren proiektu gehienek berrerabilgarritasuna dute amesturiko jomuga nagusia. Calzolari-k dioenari (Calzolari, 1989) jarraituz, bi eratara uler daiteke berrerabilgarritasuna:

- Batean ideia da, jada dauden eta eskura daitezkeen baliabide lexikalak berriro erabil daitezkeela, jatorriz pentsaturik ez zeuden helburu edo aplikazioetarako.
- Etorkizunean hainbat aplikaziotarako baliagarriak izango diren baliabideak sortu beharra.

Berrerabilgarritasunari arreta jartzeko arrazoi asko daude: ekonomikoak, estrategikoak, teknikoak, linguistika teorikoarenak zein konputazionalarenak. Halaxe dio behintzat EUROTRA-7 ikerketako koordinatzaileak (Heid, 1991). Aipatu ikerketan aplikazio konputazionaletarako baliabide lexikal eta terminologikoen berrerabilgarritasuna izango dute aztergai hamaika instituziok (industria eta mundu akademikokoak).

Iturri lexikalen berrerabilgarritasunaz hitz egiterakoan, euskarri magnetikoan dauden hiztegiez (MRDez) arituko gara. Batez ere, euskarri magnetikoan dauden hiztegiak erabili izan baitira oinarrizko informazio-iturri gisa, LNPrako lexikoiak hornitzeko zeregin horretan (Boguraev eta Briscoe, 1989; Castellón, 1993), horietan gorderik dagoen tradizio lexikografikoaz aprobetxatu nahirik. LNPko ikertalde askok jardun du MRDez baliatzen, joan den azken hamarkadan. Hala ere, MRDetatik informazioa eskuratzea ez da nolanahiko lana, teknika eta metodologia berrien garapena eskatzen baitu (LNPtik zein informatikaren

aldetik). Euskarri magnetikoan gorderik egon arren, ez da ahaztu behar giza erabiltzaileari begira eginiko hiztegiak direla, eta paperean inprimaturik daudenen arazo berberak dituztela. Arazo horiek informazioaren parte handi bat formalizatu gabe izatetik datoz. Horregatik, giza erabiltzailea bere ezagumendu linguistikoaz eta inteligentziaz baliatu beharko da MRD nahiz hiztegi arrunt gehienetako informazioaz jabetzeko.

Bestalde, formalizazio-ezaz gain (batez ere LNPrako egokia ez den heinean), MRDan aurki daitezkeen zenbait bertsiok hainbat inkontsistentzia eta akats dituzte, horietan informatika tresna lagungarri soila izan baita. Hau dela eta, tratamendu konputazionala bideratzeko aurreprozesua eskatuko dute. Hori guztia kontuan izanik, proiektu bati ekin aurretik, iturrien azterketa enpirikoa egin beharko da. Horrela diote Boguraev eta Briscoe-k (1989:35):

> *"It´s clear that very careful empirical analysis of a dictionary source must be carried out prior to any serious project..."*

Azterketa horretan, bereziki bi ezaugarri hauetaz arduratu beharko gara:

- MRDetan aurki daitekeen informazioak LNPrako duen aplikagarritasuna.
- Informazioa nola dagoen egituratua eta antolatua, eta, jakina, nola eskuratuko den MRDtik.

Dena den, arazoak arazo, eta batzuk corpusak aztertzearen aldekoak izan arren (beste batzuen artean, Zernik, 1991; Grishman eta Sterling, 1992), MRDak hartu izan dira nagusiki iturri lexikal aberatsentzat, halaxe diote behintzat Donald Walker-ek eta Antonio Zampolli-k *Computational Lexicography for Natural Language Processing* liburuaren sarreran (Boguraev eta Briscoe, 1989:xiv):

> *"The various kinds of existing dictionaries, and in particular the dictionaries available in machine-readable form, are obviously the richest and most valuable sources, based as they are on a long lexicographical tradition which encompass a treasure store of data, information and knowledge".*

MRDetako edukia automatikoki analizatu eta arakatzeko hainbat metodologia eta tresna garatu izan dira. Horrela, lexikografia konputazionalaren alorrean badago hiztegien analisirako *parser* orokorrak garatzeko xedea ere (ikus Neff eta Boguraev, 1989, 1991). MRDak erabiliz LNPko sistemetarako osagai lexikalak eraikitzeko teknikak eta metodologiak garatzea helburu dutenen artean Europako Erkidegoko ACQUILEX[27] (Esprit BRA-3030: *Acquisition of Lexical Knowledge for Natural Language Processing Systems*, (Calzolari, 1990b)) azpimarratuko genuke.

Informazio-eskuratze hori ez da behin ere osoa izango, ezta hutsik gabekoa edo erabat automatikoa ere. Jakina, MRDek soilik ezin diote LNPko behar guztiei erantzun. Gaur egunean, MRDez gain corpusa bilakatu da LNPko sistemetarako lexikoa eskuratzeko iturburu nagusietariko bat. Testu-masa handietarik informazio lexikala eskuratu nahian hainbat proiektu garatzen ari dira. Esate baterako, EDR proiektu japoniarrek itzulpen automatikorako hiztegi elektronikoak landu dituzte, eta corpusa hartu dute iturri nagusitzat horiek eraikitzeko eta aberasteko.

---

27 ACQUILEX: Esprit BRA 3030, batez ere, MRDetan oinarrituko da. ACQUILEX II: Esprit BRA 7315, bigarren honetan arreta handiagoa jarriko dute corpusean. Hala ere, biek zuten helburu nagusi bera: jada existitzen diren baliabideen berrerabilgarritasuna aztertzea LNPrako sistemak eraikitzeko.

Esan gabe doa lexikografoen lana ere ezinbestekoa dugula lexikoiak osatzeko, MRDen gabeziak eta corpus-lanketarako tresnen zailtasunak kontuan izaki. Haien lana errazteko ingurune lexikografikoen garapena ere oso inportantea dugu, gaur egun lexikografoek laguntza informatiko handia eskura dezakete: KWIC delako konkordantzi programak, informazio-eskurapenerako sistemak linean, *lexicographer's workbench* delakoa (Lenders, 1990) , eta abar dira, besteak beste, zenbait kasu.

Aipaturiko hiru iturriok, hala nola, MRD, corpusa eta lexikografoen lana elkarren osagarri dira LNPko baliabide lexikalen premiari erantzuteko.

Hala ere, izan dira ikerkuntzarako beste joera batzuk ere lexikoi konputazionalak eraikitzeko. Hona hemen zenbait: ezaugarri-egiturak, hautapen-murriztapenak, eta azpikategorizazioa esplizitu egiten dituztenak sintaxian oinarrituz (Gross, 1975); hitzei buruzko mota guztietako informazioa modu formalizatuan jasotzen duten hiztegiak (Mel´cuk eta beste, 1981), lexikoi konputazionalen diseinua LNPko sistema desberdinetarako (Flickinger eta beste, 1985).

Bukatzeko, euskarazko iturriei gagozkiola, hor ditugu I. Sarasolaren *Euskal Hiztegia*, UZEIren *Sinonimoen Hiztegia* eta *Atzekoz aurrera*, eta Elhuyarrek, Harluxet Fundazioak eta Adorez taldeak, besteak beste, kaleratutako hiztegi-lanak euskarri elektronikoan.

8.4.3 Informazio lexikalaren errepresentazioa

Datu lexikalen eskurapenaren arazoari begira "berrerabilgarritasuna" hitz gakoa bada ere, gaur egun esan genezake interesak errepresentazioaren aldera lerratu direla. Lexikoaren premia duten guztien lan bateratu eta koordinatu baten beharra nabari da. Horrela, bada, estandarizazioa da egin beharreko ezinbesteko urratsa. Proposamenak hasi ziren plazaratzen 1990. urtean (Sperberg-McQueen eta Burnard, 1990), hiztegien errepresentaziorako eskemek orokorrak eta aplikazioetarik independente izan behar dute. Behar horri erantzuteko asmoz, elkarlanerako bideak aurkitu eta informazio-trukea bermatuko duten ekimenak sustatuko dira; beste batzuen artean, honako hauek aipatuko genituzke: *Text Encoding Initiative (TEI)*[28] (Sperberg-Mcqueen eta Burnard, 1994)*, The ACL Data Collection Initiative, Consortium for Lexical Research.*

Lexikografia konputazionalaren alorrean lexiko-sistemen azterketa, errepresentazioa eta erabilera, gero eta garrantzi handiagoa hartzen ari dira. Azken hamarkadan lexikoigintzan aurrera egin da: erredundantziaren arazoa konponduz, datuen kontrola eta kontsistentzia gauzatuz, eta informazio-atzipena erraztuz.

Orainokoan nabarmendu dugun "benetako" lexikoien premia horrezaz gain, landurikoetan ez zegoen adostasunik ez lexikoiek jaso behar zuten informazioaz ez hori nola errepresentatu behar zen (Ingria, 1986). Gai honen haritik, informazioa eskuratzeko prozesuak ez ezik, errepresentazioaren arazoak ere ernatzen ditu, batik bat, ikertzaileen interesak. Kontuan izan behar da, 1.3 atalean esan bezala, HKren egungo joeraren arabera, hizkuntza-ezagutza gramatikaren arlotik lexikoarenera lerratu dela, eta

---

28 TEIren helburua, dena den, ez da lexikoaren arlora mugatzen. Ekimen horretako gidalerroen helburua giza zientzietako ikerkuntzan datu-trukerako eta testuen kodeketarako formatu estandarra eskaintzea da.

ikusmolde-aldaketa horrek gramatikak erraztea ekarri duela. Baina informazioa lexikoan pilatzeak sarrera lexikalak informazio erredundanteaz hornitzea ekar lezake. Informazioaren kopuruak eta konplexutasunak informazioa bera kontrolatzeko arazoak sor ditzake. Beraz, beharrezkoa izango da sarrera lexikalek zein motatako informazioa behar duten erabakitzeaz gain, nola egituratu informazio hori guztia erredundantzia ekiditeko eta portaera bereko hitz moten arteko pareko ezaugarriak antzemateko. Arazo horiei erantzuteko ezagutza-base lexikalak (*Lexical Knowledge Bases*, LKB) garatuko dituzte zenbait proiektutan, adibidez ACQUILEXen. Ezagutza-base lexikalek, herentzia-mekanismoak eta erregela lexikalak baliatuz, informazio lexikalaren erredundantzia ekiditea eta kontsistentzia bermatzea lortzen dute. Horretaz gain, informazio lexikal egituratua errepresentatzeko orduan, ahalik eta zehatzen izateko eta orokor diren tasunak jasotzeko, batez ere, herentzia, balio lehenetsien espezifikazioa eta erregela lexikalak aipatu izan dira. Tresna horien azpian dagoen ideia da hitz moten hierarkia eta herentziaren nozioa. Hau da, hitz mota bereko elementuek ezaugarri berak konpartituko dituzte. Esate baterako, erregela lexikalen zeregina izango litzateke bi hitz motako elementuen arteko harremanak sistematikoki errepresentatzea (Flickinger, 1987). Ildo beretik, semantika lexikalaren ikuspegitik, item lexikalak errepresentatzeko *Qualia Structure* teoria garatzen du Pustejovsky-k (Pustejovsky, 1991). Teoria horren bidez, hitzek dakarten polisemia sistematikoki adieraziko da lexikoian behar ez den anbiguotasun lexikala ekidinez. Horrez gain, autore horrek dio, egitura lexikal bakanak ezagutza-base lexikal zabalago batean integra daitezkeela herentzia lexikalaren teoriari esker. Teoria horrek lexikoiaren antolamendu orokorrerako behar diren printzipioak ditu, eta gure hizkuntza naturalaren lexikoia osotasun kontzeptual batean integratzen laguntzen digu.

Bestalde, hiztegi-informazio lexikalaren egitura konplexua errepresentatzea oso zaila izan daiteke. Datu-eredu "konbentzionalak" desegokiak dira datu-base lexikaletarako. Adibidez, erlazionalean informazio lexikalaren egitura konplexua ezin da egoki errepresentatu. Egokiagoak bide dira datu lexikalak errepresentatzeko ezaugarri-egituretan oinarritutakoak, besteak beste arrazoi hauengatik:

- Informazioa atzitzeko eta maneiatzeko bide anitz.
- Hiztegi jakin baten antolaketa gordetzen ahal da, kontsultarako "transparente" eginez.
- Oinarri teoriko sendoa.
- Lexikoi konputazionalekiko bateragarritasuna.

Ezaugarri-egituretan oinarritutako errepresentazio-eredua inplementatzeko modurik egokientzat-edo, objektuei zuzenduriko datu-basea daukate (Ide, N. eta beste, 1993).

Datu-base lexikal (DBL) idealaren ezaugarriak hauek behar lukete izan:

- Erabiltzailearen eta datu-basearen arteko elkarrekintza oso garrantzitsua da. Hori gauzatzeko komeniko litzateke interfaze lagunkoiak izatea.
- Malgutasuna, hau da, edozein mementotan helburu berrietarako egokitzen erraza, informazio mota berriak aise onartuko dituena.
- Berrerabilgarritasuna, hots, informazio lexikala berrerabilgarria izatea.

- Dimentsio-aniztasuna, hau da, helburu askotarakoa. LNPren arloko aplikazioetarako zein lexikoa beharrezkoa den beste aplikazio batzuetarako ere baliagarria. Horretarako informazio askotarikoa beharko du izan: morfologikoa, sintaktikoa, semantikoa, pragmatikoa, etab.
- Neutraltasuna (eskola linguistiko desberdinekikoa), hau da, bertan egindako deskribapen linguistikoak ez lituzke baldintzatu behar etorkizuneko aplikazioak.

Garbi dago, beraz, datu lexikalak, hiztegietakoak esaterako, oso datu konplexuak direla eredu konbentzionalari jarraitzen dioten datu-baseen bidez errepresentatzeko. Horregatik, zenbait autore ezaugarri-egituretan oinarritzen diren ereduen aldekoak dira. Horrez gain, nabarmena da zein garrantzitsua den datu lexikalak datu-base lexikal batean gordetzea; besteak beste, datu-baseek eskaintzen dituzten aukerei esker hiztegien eguneratzea eta mantenimendua, hiztegien berstio desberdinen sorkuntza, datuen kontsitentzia bermatzea, etab. oso modu ziurrean egin daitezkeelako.

Aurreko ataletan nabarmendu nahi izan dugu berrerabilgarritasun- eta estandarizazio-asmoek eragin handia izan dutela lexikoa nola adierazi erabakitzean.

Atal honen bukaeran, labur-labur bada ere, informazio lexikalaren adierazpide nagusiak definituko ditugu:

Hiztegi datu-baseak

Demagun hiztegi oro, euskarri elektronikoan badago, datu-basetzat har daitekeela. Beraz, hiztegi datu-baseen definizio sinple-sinplea hartu dugu abiapuntu. Hiztegiak biltegiratzeko moduak hiru sailetan bereiziko ditugu:

Testuak

Oro har, testu soilak dira, egitura nabarmenik izan ez eta atzitzeko erraztasunik ematen ez dutenak. Baieztapen orokor horrek eskatzen du, ordea, nolabait ñabartzea. Testu-fitxategi izanagatik, gutxi dira egituratze tipografikoa-edo agertzen ez dutenak (betiere ustiatze zailekoak).

Datu-base erlazionalak

Errepresentazio-eredu erlazionalean objektuak eta entitateak identifikatzen dira lehenik, gero beren ezaugarriak zehaztu eta azkenik elkarren arteko erlazioak adierazi. Eredu honetan hiztegiak errepresentatzea eginkizun zaila da. Nolanahi ere, atzibide hobeak eskaintzen dituzte testuzko formatuek baino.

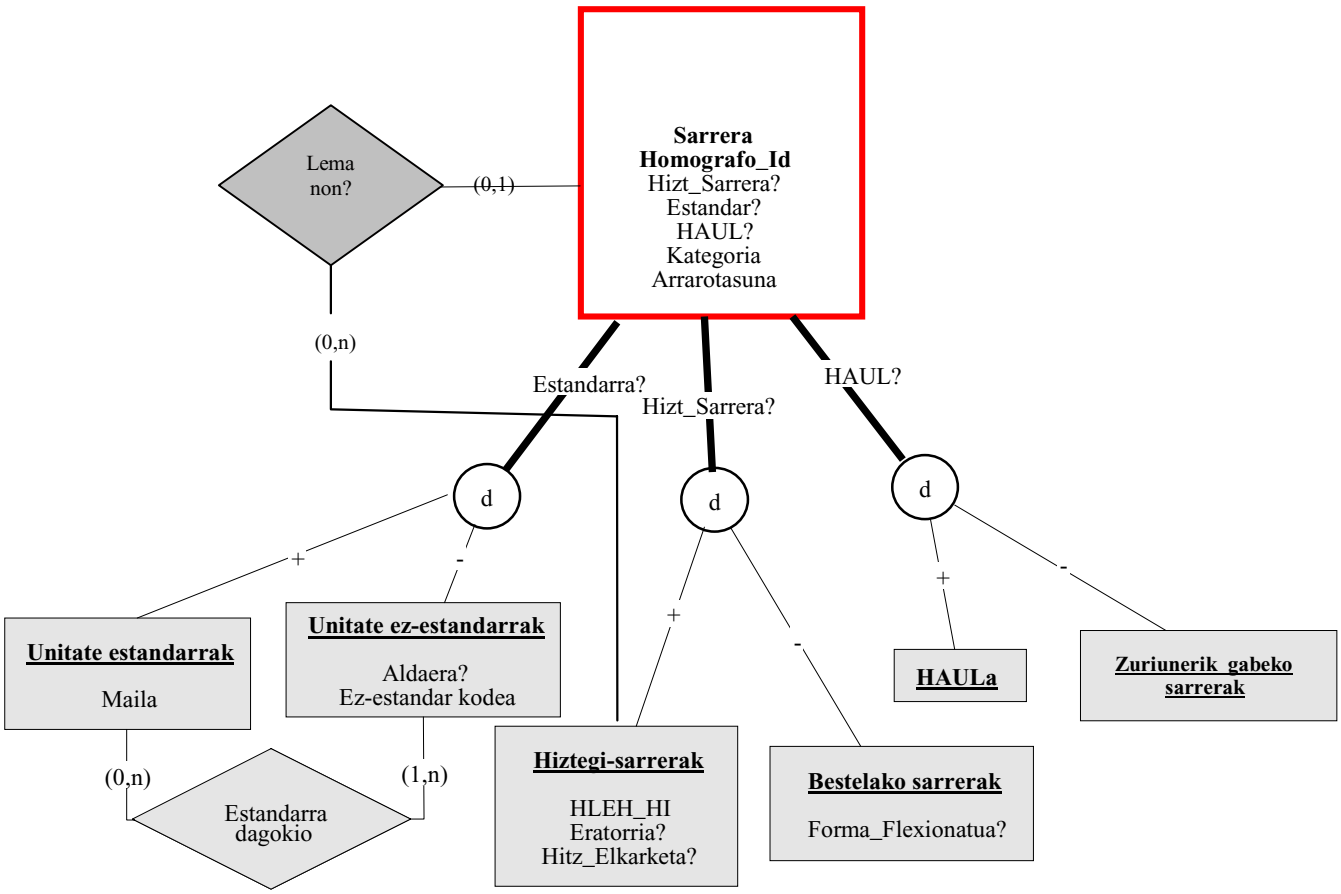
Testu etiketatuak (SGML, XML...)

Testu-fitxategiak dira, baina testuaren egitura islatzeko markez hornituta daude. Eredu erlazionala baino malguagoa da. Markatzea deskribatzailea da (ez prozedurala).

Datu-base lexikalak

Datu-basea informazioa modu egituratuan gordetzeko erabiltzen den baliabide konputazionala da; hau da, informazio-biltegi erraldoi bat, ezagutzen ditugun gordailu fisikoen pareko (artxibategiak, biltegiak...).

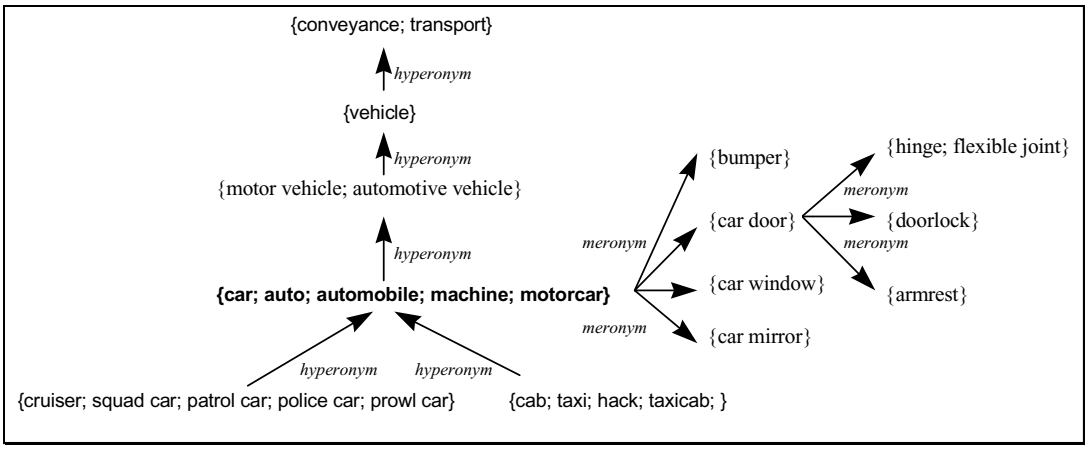
Euskararen datu-base lexikala (EDBL), euskararen prozesamendu automatikorako baliatzen den informazio lexikalaren biltegi nagusia dugu. Besterik gabe, datu-base lexikalaren zer-nolakoak ilustratzeko eskema orokorra ekarri dugu hona (ikus 45. irudia).



45. irudia. EDBLko unitateak eta goi-mailako espezializazioak

Ezagutza-base lexikalak

Semantika lexikala modu egituratuan biltegiratzeko euskarriak dira ezagutza-base lexikalak. Ikus dezagun adibidez, WordNet 1.5 bertsioko adibide bat:



WordNet oinarri hartuta, EuroWordNet garatu da Europako zenbait hizkuntzatara hedatuz, horien artean, euskarara ere. Ikus dezagun adibide bat:

EuroWordNet-eko adibidea:

WN 1.5 :	*tree_1*	"plant"	**07991027n**		*tree_2*
"diagram"	**08514899n**	***SpanishWN:***	*árbol_2*	"planta"	
07991027n		*árbol_3*	"diagrama"	**08514899n**	
EusWN:	*zuhaitz_2*	"landarea"	**07991027n**		
zuhaitz_3	"diagrama"	**08514899n**			

8.5 Corpusen linguistika

Baliabide lexikalen artean aipatu dugu corpusa. Eta bereziki azpimarratu nahi dugu, geroz eta garrantzi handiagoa hartzearen ondorioz, alor bihurtzeko bidean baita. Izatez, egun, *corpusen linguistika*z hitz egiten da, corpusen inguruan egiten den lan eta ikerkuntzari erreferentzia egiteko.

Errealitateari buruzko hipotesia bota nahi duen zientzia orok, errealitateak ematen dizkion datuetan oinarritu beharko ditu haren ondorioak. Datu horiei natur zientzietan datu enpiriko deitu ohi zaie. Giza zientziek ere datu enpiriko horietara jo dezakete deskribatzen duten errealitateari buruzko hipotesiak ziurtatzeko. Baina hizkuntzalaritzaren kasuan zeintzuk lirateke erreferentetzat hartu beharreko datu enpiriko horiek? Berez, hizkuntza izango da hizkuntzalariak deskribatu nahiko duena, baina hori egiteko enuntziatu linguistikoetara jo beharko du, hau da, hizkuntzaren erabilera gauzatzen duten ekintzetara, bai ahozko hizketara, bai eta idatzizko testuetara ere. Bestela esanda, hizkuntzalaritzak ere haren teoriak babestuko dituen eta hizkuntzaren joera nagusiak erakutsiko dizkion erreferentzia-elementuak (datu

enpirikoak) beharko ditu. Corpusek, erreferentzia hori sistematizatuko duten testu edo hizketa multzoa osatzen dute. **Hizkuntzalaritza enpirikoa** deritzona izan zen mende hasieran zegoen korronteetako bat.

Hasierako hizkuntzalaritza sortzailearen (Chomsky, 1957) oinarrian, ordea, hizkuntza batek izan ditzakeen enuntziatuak ezin zenbatuzkoak direla zegoen, eta horien ustez, ez dago hizkuntzaren mekanismoak azalduko dituen eta datu egokiak dituen testu multzo (corpus) finiturik. Deskribatu behar den objektuaren adibidea bere hizkuntzaz hitz egiteko konpetentzia duen hiztun ideal batengan bilatu behar dela diote. Korronte haren ondorioz, hizkuntzalaritzaren azterketa, ikuspegi enpirista batetik ikuspegi **arrazionalista** batera igaro zen. Orientazio berri honi egiten zitzaion kritika zen hizkuntza deskribatzeko corpusek ez zutela baliorik.

Aurrerago, eta batez ere hizkuntzalaritza aplikatua egiten hasi zenetik, corpusen helburua hizkuntzaren ikuspegi osoa ematea baino, beste era batera jardutea izan da. Corpusen helburu berria da hizkuntzalaritzaren ikerkuntzan oinarri izango den lagin eredugarri bat izatea, datu objektiboak eskainiz. Corpusa ezingo da hizkuntzarekin parekatu; ezaugarri egokiak edo ez hain egokiak izango dituen datu multzoa izango da. Helburu berri horren harira hizkuntzalaritza enpirikoa eta, berarekin batera, corpusen gaineko interesa handitu egin da berriz. Hizkuntzalaritza arrazionalistaren eta enpirikoaren arteko eztabaidak bere horretan dirauen arren, corpusen erabilgarritasuna ez da egun zalantzan jartzen.

Corpusa, oro har eta zehaztasunetan sartu gabe, hizkuntzari buruzko datu-bilduma da. Adibidez, lexikoaren morfologian interesatuta dagoen hizkuntzalari batek corpusa hizkuntza bateko hitz eratorrien multzotzat har dezake; sintaxiarekin lan egiten duen hizkuntzalari batek, berriz, hizkuntzaren sintagma multzo zabaltzat.

Baina hori guztia hala izanik ere, *corpus* hitza modu zorrotzago batean erabiltzen dela esan daiteke, eta oso lotuta dago *hizkuntzaren teknologiak* deritzon arloari. Corpusak sortzeko teknikek, beraiez baliatzen diren sistemek, beraien informazio mota ezartzen duten irizpideek, horiek guztiek egiten dute corpusen sorrera eta erabilerak oso leku garrantzitsua izatea hizkuntzaren prozesamenduan, *baliabide linguistiko*en eremuan, hain zuzen.

8.5.1 Corpusaren ezaugarriak

Testu bat baino gehiago duten bildumak *corpus* dei ditzakegu –corpus latineko hitza dugu ‘gorputza’ adierazteko, eta beraz, corpusa edozein testu-gorputz izango da–. Baina *corpus* hitza hizkuntzalaritza modernoan erabilitako adieran, definizio xume horrek baino ezaugarri zehatzagoak dituela esan dezakegu. Hona corpusen ezaugarri nagusiak:

- Eredugarria eta adierazgarria

  Zenbaitetan hizkuntzalaritzan ez zaigu interesatzen egile baten testu hau edo bestea, baizik eta nolakoa den hizkuntza baten aldaera, edo zein diren aldaera jakin baten ezaugarriak. Honelakoetan bi aukera ditugu datuak jasotzeko garaian:

  - Aldaera horrek izan dezaken agerpen bakoitza analiza dezakegu –nahiz eta aukera hau ezin den aurrera eraman, zenbait kasutan izan ezik; adibidez, hildako hizkuntza batekin

zeinek testu gutxi batzuk izango dituen–. Hala ere, gehienetan kasu bakoitza aztertzea amairik gabekoa izan daiteke, eta ezinezkoa.

- Aldaera horren adibidetegi (lagin) txiki bat sor dezakegu. Hau aurrekoa baino errealagoa eta errazagoa izango da.

Ondorioz, saiatu behar dugu aldaera horri dagokion corpusik esanguratsuena eraikitzen. Eta horretarako honelakoa izan beharko du: corpusak aldaera horren tendentzien eta batez bestekoen berri eman beharko du, eta ahal den modurik zehatzenean, gainera. Bilatzen duguna egile eta genero sorta ahal den zabalena biltzea da, eta horiek elkarrekin jarri ondoren, aztertzen ari garen aldaeraren informazio orekatua eta zehatza emateko egoki izango dena.

- Neurri mugatukoa

Corpus elektronikoek milioi bat hitz baino gehiago izan ohi dituzte, baina nolako luzera dute milioi bat hitzek modu ulergarriago batean esanda?

Aldizkari bateko orrialde beteko artikuluak 965 hitz izan ditzake (*New Yorker*). Aldizkariak 112 orrialde izango balitu, eta orrialde guztiek hitz kopuru berdina izango balute… milioi bat hitz 9 ale lirateke.

Liburu baten orrialde batek 374 hitz baldin baditu (*English Corpus Lingistics,* 3. orria), eta liburuak 338 orrialde baldin baditu, liburu osoko orrialdeek hitz kopuru berdina izango balute, 126.000 hitz izango genituzke. Beraz, milioi hitz izateko honelako 8 liburu beharko genituzke.

Axularren *Gero* liburuak (1634) 100.000 hitz inguru ditu (247 DIN-A4 orrialde). Etxepareren *Linguae Vasconum Primitiae* liburuak (1545) 7.000 hitz inguru ditu (37 DIN-A4 orrialde, poesia).

Pasaiako Hizkera liburuak (266 orrialde txiki) 40.141 hitz ditu. Honelako 25 liburu beharko genituzke milioi bateko corpusa osatzeko.

Corpusak gehienetan *estatikoak* dira. Hots, behin neurri bat hartuta, bere horretan uzten dira edozein azterketarako erreferentzia modura. Gutxiagotan corpusak handituz eta aberastuz doaz, datu berriekin eta analisi mota berriekin. Azken horiek, ordea, ez dira iturburu fidagarria datu kuantitatiboei dagokienez (datu kualitatiboekin ez bezala). Horien neurria behin eta berriz aldatuz doa, eta ez dira corpus mugatuak bezain eredu zehatzak. Horregatik joaera da corpus estatikoak egitea.

- Makinan irakurtzeko moduan jarria (MRD)

Gaur egun *corpus* hitzak berekin dakar konputagailuan irakurtzeko moduko izatearen ezaugarria. Hau garai batean ez zen horrela; izan ere, lehen *corpus* hitzak inprimatutako testua adierazten baitzuen.

Makinan irakurgarri diren corpusek idatzizko edo ahozkoen ondoan honoko abantailak dituzte:

- Bilaketa eta erabilera oso azkarrak ahalbidetzen dituzte.
- Informazio berriz berehala osatu ahal dira.

- Erreferentzia estandarra

8.5.2 Corpus motak eta beren erabilgarritasuna

Bi corpus mota bereizten dira: ahozkoa eta idatzizkoa.

Ahozko corpusen definizioa, helburuak eta mamia aldatu egiten dira zein tradiziotan kokatzen garen. Oro har, bi tradizio daude: batetik, *corpusen linguistika*ren tradizioa eta bestetik, *fonetika eta ahozko tratamendua*rena.

*Corpusen linguistika*ren tradizioan, ahozko corpus bat ahozko aldaera baten transkripzio ortografikoa da. Transkripzio hau aberastu egiten da gure helburuen arabera beharrezko izan ditzakegun elementuekin. Azken finean, corpus hau ahozko erabileraren errepresentazio sinbolikoa da, eta, oro har, estilo, erregistro, edo komunitate baten errepresentazioa da.

*Fonetika eta ahozko tratamendua*ren tradizioan, berriz, ahozko corpusaren ezaugarririk garrantzitsuena ahozkotasunaren grabaketa da, helburua informazio fonetikoa lortzea baita, eta horrela sintesiaren eta ezagutzaren inguruko tresnak garatu ahal izatea. Errepresentazio sinbolikoa alfabeto fonetiko baten bitartez egin ohi da, nahiz eta, zenbaitetan, erabileraren izenean, errepresentazio ortografikoa ere egin ohi den.

Nolanahi ere, ahozko corpusak behar-beharrezkoak dira ondoko lan-alorretan:

- Fonetika-lanetan, hizkuntzen deskripzio segmentala eta suprasegmentala egiteko, bai akustikari begira eta bai ebakidurari begira.
- Fonetika aplikatuak dituen hainbat adarretan: psikolinguistikan, hizketaren sorkuntzan eta jabekuntzan, interferentzia fonetikoan eta bigarren hizkuntzen jabekuntzan, hizketaren patologiak aztertzerakoan.
- Fonetikan oinarritutako soziolinguistika eta dialektologiaren ikerketetan, testua fonetikoki transkribatuta eta hainbat mailatan kodifikatuta izateko.
- Ahotsaren tratamenduan, testua ahots bihurtuko duten aplikazioak garatzeko. Aplikazio horien artean ditugu: sintesi-unitateen hiztegia osatzen duten unitate fonetikoen erauzketa; unitate horien arteko loturaren ikasketa; kontu prosodikoen datuen eskurapena; soinuen iraupenen berri edo esaldien intonazioari buruzko datuen eskurapena; ezagutza-unitateen modelo akustikoak sortzea; hiztunak eta ingurunea sailkatuko dituen datuak lortu ahal izatea; hizkuntza-ereduak sortu ahal izatea; eta ezagutzaileek erabiliko duten lexikoa osatu ahal izatea.
- Hiztunaren identifikazio eta ezagutzarako, zenbait pertsonaren hitz egiteko moduak bilduz.
- Elkarrizketa-sistemak garatzeko; pertsonen arteko elkarrekintza nahiz makina eta pertsonen artekoa jasotzen duten testuen bidez, erabiltzaileek egiten dituzten galdera motak ezagutu ahal izango ditu sistemak.
- Diskurtsoaren eta elkarrizketaren analisian.
- Analisi gramatikalean, idatzitako hizkuntzan lortutako datuekin parekatzeko.

Idatzizko corpusak, berriz, idazleen eskutik sortutako testuak dira, narrazio, poesia, saiakera… modura ekoitziak. Idatzizko testuak behar-beharrezkoak dira beroietan oinarrituta hizkuntzaren alor guztiak aztertzeko. Idatzizko corpusen bidez, hizkuntzaren atalei buruzko informazioa eskuratu, egiaztatu eta formalizatu ahal da. Idatzizko corpusak —*gordin* nahiz *mailatan etiketatuta*— behar-beharrezkoak dira arau bihurtuko denak erabilera aldetik bermea duen egiaztatzeko, eta aldi berean arautik urruntzen dena identifikatzeko. Laburki, ondoko lan-alorretan dira oso erabilgarriak idatzizko testuak:

- Lexikografian. Datu enpirikoak aspaldi erabili izan dira lexikografian, baita corpus linguistikoen arloa sortu baino lehen ere. Samuel Johnson-ek, kasu, bere hiztegia literaturako adibideez hornitu zuen, eta XIX. mendean, *Oxford* hiztegiak, aipuak fitxetan gordetzen zituen erabilera azaltzeko. Corpusek, hala ere, lexikografoen hizkuntza ikusteko modua aldatu dute. Hizkuntzalari batek makinan irakurtzeko moduko corpus bat esku artean duenean galdeketak ditzake, eta segundo gutxitan hitz baten edo sintagma baten adibide asko eta asko eskura ditzake. Horrela, hiztegiak lehen baino askoz azkarrago sor eta errebisa daitezke, eta horrek hizkuntzaren informazioa beti guztiz eguneratuta edukitzeko aukera ematen digu. Gainera, definizioak osatuagoak eta zehatzagoak dira adibide asko daudelako. Ondoren azalduko dugun adibidea Atkins eta Levin-etik (1995) hartuta dago. Ingelesezko *shake* motako aditzak aztertu zituzten, eta haien definizioak hiztegi hauetatik hartu:

    - *The Longman Dictionary of Contempary English*
    - *The Oxford Advanced Learner's Dictionary*
    - *The Collins COBUILD Dictionary*

  Shake motako aditzen artean *quake* eta *quiver* ziren. Bi hiztegik, *Longman* eta *COBUILD*-ek, bi aditz horiek iragangaiztzat jotzen zituzten, baina *Oxford* hiztegiak *quake* iragangaiztzat eta *quiver* iragankortzat. Autoreek hitz horien agerpenak bilatu zituzten corpus batean (50.000 hitz). Corpuseko adibideetan *quake* aditzak erabilera iragankorra ere bazuela ikusi zuten. Adibidez: *It quaked her bowels* eta *quivering its wings*. Bestela esanda, horrekin erakutsi zuten hiztegiek ez zutela informazio erabat zehatza ematen, bi aditzak iragangaitz eta iragankor izan baitaitezke. Ikus dezakegu corpus errepresentatibo batek indar dezakeela edo bertan behera bota lexikografoaren intuizioa.

  Lexikografiari lotuta, hiztegien sorkuntzan eman dezake laguntza handia corpusak. Corpusak erabilita, lexikografoek jakin ahal izango dute ea hitz berririk sartu diren hizkuntzan edo lehendik dauden hitz zenbaitek haien esanahia aldatu ote duten, eta abar.

  Bestetik, corpusetatik ateratako adibideak analisi-taldeak egiteko moduan antola daitezke. Adibidez, hitz baten eskuinean agertzen den testua kontuan hartuz gero, testu hau alfabetikoki antolatu ahal izango dugu, eta kolokazio jakin baten berri emango digu, denak elkarren segidan agertuko direlako. Gainera, corpus-datuek informazio testual handia dutenez –lurralde bateko aldaera, egilea, data, generoa, kategoria mailako etiketak…–, errazagoa izango da lotzea zenbait erabilera zonalde jakin bateko erabilerarekin, edo aldaera jakin batekin, edo genero jakin batekin.

Beraz, hitz soilak baino, hitz multzoak sortzeko gaitasuna, eta elkarrekin agertzen diren hitzen arteko erlazioa aztertzeko emaitzak lor daitezke corpusetatik.

- Gramatikan. Corpusak oso erabilgarriak dira ikerketa sintaktikoak gauzatzeko, honokoengatik:
    - Hizkuntza baten aldaera jakin bat ordezkatzeko datu kuantitatiboak izango ditu, balizko mailan, bederen.
    - Datu enpirikoak bertan izango ditugu gramatika-teorietatik sortutako hipotesiak probatzeko (eta baieztatzeko edo ezeztatzeko).

  Nijmegen Unibertsitatean, kasu, oinarrizko gramatika formalak eguneroko hizkuntzarekin probatzen saiatzen dira corpus hauen bitartez (Aarts & Meis, 1986). Gramatika formala lehenik tresna batera bihurtzen dute. Ondoren, analizatzaile sintaktiko batera bidaltzen dute, eta ondoren corpusaren aurka lan egiten da. Ondoren, gramatika aldatuz doa egindako analisien okerrak ikusteaz batera.

  Bestetik, kategoria lausoak eta graduazio-kontuak zehazten ditu corpusak. Hizkuntzalaritza teorikoan, kategoriak "azkar eta pisu" moduan ikusi izan dira –termino bat kategoria batekoa izan edo ez–. Hala ere, kategoria mailan egindako lan psikologikoek diotenez, kategoria kognitiboak ez dira "pisu eta azkarrak" izaten, baizik eta muga lausodunak (*fuzzy*), beraz galdera ez da ea termino bat kategoria batekoa den ala ez, baizik eta zenbatetan den kategoria batekoa besteari aurre eginez. Corpusetan agertzen den lengoaia naturalari begiratuta, argi dago kategoria lauso hauek hobeto datozkiela datuei: muga zehatzak ez dira existitzen; nahiz eta zenbaitetan badiren halako graduatzaileak.
- Semantikan. Corpus linguistikoek semantika laguntzen dute hurbilpen neutral bat egiteko garaian. Mindt-ek (1991) azaldu zuen nola erabil daitekeen corpus bat termino linguistikoei esanahiak emanez irizpide objektibo batzuekin. Mindt-ek zioen, semantikan, gehienetan terminoei ematen zaizkien esanahiak gure intuizioetan oinarritutakoak direla. Mindt-ek dio aldaketa semantikoak testuetan nabari daitezkeela testuinguru-aldaketei esker (sintaktikoak, morfologikoak eta prosodikoak), eta kontuan hartuz entitate linguistiko horien testuingurua, lor dezakegu ezaugarri semantiko objektiboa enpirismoan oinarrituta. Corpus berezituetatik (hiztegietatik, kasu) hitzen arteko erlazio lexiko-semantikoak atera ahal izango ditugu. Hau oso garrantzitsua izan daiteke, adibidez, horri esker inguruan dituen hitzen arabera jakin ahal izango dugu hitz baten adiera corpusean. Izan ere, inguruko hitzekin zein erlazio mota duen ikusita, desanbiguatu ahal izango baitugu.
- Pragmatika eta diskurtsoaren analisian. Corpusa gutxi erabili da oraindaino alor hauetan. Izan ere, testuingurua oso garrantzitsua da berauetan (Myers, 1991), eta corpusetan erabiltzen diren testuen adibideak beraien testuingurutik atereak egoten dira. Zenbaitetan garrantzitsua den informazioa (generoa, klasea, erlijioa) kodetzen da corpusaren barruan, baina ezinezkoa izaten da informazio hori erauztea corpusetatik. Hala ere, Marcu-k (2000) erakutsi du posible dela testu baten egitura diskurtsiboa adieraztea, horretarako eskuz etiketatutako corpus batean oinarritutako algoritmoa erabiliaz. Gainera, hainbat aplikaziotan lagungarria izan daitekeela aldarrikatzen du,

hala nola, laburpenak egiteko, itzulpen automatikorako, etab. Marcu-ren lanaren ondoren, corpusetan oinarritutako diskurtsoaren azterketak bultzada ederra jaso dezake.

Idatzizko corpusak orain arte zeregin lexikografiko, gramatikal eta semantikorako erabili izan dira batez ere. Egun, hizketaren tratamenduak garrantzia hartu duen heinean, ahotsaren tratamenduari begira, alderdi fonologikoak ikertzeko oinarri bihurtu da. Bai eta pragmatikari loturiko alderdi batzuk aztertzeko ere.

Informazioaren gizartean hizkuntza batek duen garrantzia, aplikazioak garatzeko dituen baliabide linguistikoen arabera neurtu ohi da. Baliabide horien artean, corpus handien garapena lehenetarikoa izan ohi da.

Europan garatutako corpusen artean lehenetariko bat *FRANTEXT* izan zen, Institut National de la Langue Française deritzonak garatua, eta gaur egun *Trésor de la Langue Française*-ren erredakzioak erabiltzen du. Gaur egun, *FRANTEXT* hau 150 milioiko corpusa da eta suskripzio bidez atzitu daiteke.[29]

Gaur egun, konputagailuek ematen dituzten erraztasunez baliatuta, eta materialen prozesamenduari esker, neurri handiko corpusak sor daitezke, gero eta prozesu automatikoagoak erabiliz; halakoa dugu *British National Corpus*-a; 4000 testu zati ditu, eta denera 100 milioi hitz ditu.[30]

Beste adibide bat *Bank of English* dugu. Hau Birmingham-eko Unibertsitatean sortutako COBUILD proiektuaren garapena da, eta 450 milioi hitz ditu.[31]

Corpusa	Hitz kopurua	Hizkuntza
FRANTEXT	150 M.	Frantsesa
British National Corpus	100 M.	Ingelesa
XX.MECE	4.658.036	Euskara
CREA	130 M.	Gaztelania
CORDE	136 M.	Gaztelania
Bank of English (COBUILD)	450 M.	Ingelesa

Gaur egun sortzen ari diren behar komertzialei eta corpusen eguneraketari aurre egiteko *monitor* corpusak sortu dira. *Monitor* corpusa etengabe eguneratzen ari den corpusa da, eta etengabe handituz doa. Corpus hauen erabilera usuena lexikografian eta hizkuntzaren prozesamendu naturalean izango da.

Gaur egun corpusen kopurua eta horiek sortzeko proiektuen kopurua gero eta handiagoa da. Web orri honetan topa daiteke corpus horien informazio nahiko eguneratua:

---

[29] http://zeus.inalf.cnrs.fr/

[30] http://www.hcu.ox.ac.uk/BNC

[31] http://titania.cobuild.collins.co.uk

http://www.ruf.rice.edu/~barlow/corpus.html

8.5.3 Kodeketa eta markaketa

Corpus bat inongo markaketarik gabe baldin badugu, *corpus gordina* dela diogu. *Markatutako corpusak*, berriz, informazio linguistikoarekin hobetuta daude. Jakina, markatutako corpus baten erabilera askoz handiagoa da. Testu markatuetan, testu batek ezkutuan duen informazioa agerian uzten da, hainbat markaketa-prozesuren bitartez kodetuz, eta gero horiek baliatzen dira hainbat zereginetarako.

Adibidez, *darama* formak duen informazio inplizitua hau dugu: "orainaldia, singularreko hirugarren pertsona, aditz trinkoa". Hori berreskuratu ahal izango dugu gure buruak euskarari buruz duen ezagutzari esker. Baina, markatutako testu batean *darama* forma honela ager daiteke, "darama ADT A1 NOR_HU NRK_HU", eta kodeketa horrek hauxe esan nahi du: ADT aditz trinkoa dela, A1 orainaldia dela, NOR_HU nor hirugarren singularrekoa izango dela, eta NRK_HU ergatiboa ere hirugarren singularra izango duela. Markaketa horrek erraztu eta azkartu egiten du corpus batetik markatutako informazioa erauztea.

Markatze-kontu hauen harira, Leech-ek (1993) testu bat markatzeko garaian 7 maxima kontuan izan behar direla dio:

1. Markatutako corpus batetik aukera izan behar dugu marka guztiak ezabatzeko eta berriro ere corpus gordinera itzultzeko. Hau zenbaitetan oso erraza izango da: adibidez, azpiko marraren ondoren agertzen den guztia kentzean "Claire_NP1 collects_VVZ shoes_NN2" beste hau bihurtuko dugu: "Claire collects shoes". Baina, adibidez *London-Lund* corpusean intonazioa markatuta dago, eta marka hau nahasian ageri da hizkien artean: adibidez, "going" hitzak lehen silaban gorantz egiten duela adierazteko "g/oing" jartzen dute. Horrelakoak topatuz gero, zail izan ohi da testu gordina berreraikitzen.
2. Testuetatik markak erauzteak erraza izan behar du. 1. maximaren beste aldea dugu hau. 1. eta 2. puntuak elkarrekin jarriz gero, markatutako corpusak ahalik eta malgutasun handiena izan behar du erabiltzailearen erabilera errazteko.
3. Markaketa-eskema azken erabiltzailearentzat erabilgarri izango diren gida-lerroetan oinarrituko da. Corpus guztiek dute eskuliburua. Bertan erabilitako markatze-eskemaren zehaztasunak eta erabilitako gida-lerro guztien azalpenak ematen dira. Horrek ahalbidetuko du erabiltzaileak berehala ulertzea marka bakoitzak zer adierazi nahi duen, eta anbiguitate-kasuetan zergatik egin ote den bataren edo bestearen alde.
4. Argi utzi behar da testu bat nola eta nork markatu duen. Corpus bat eskuz marka daiteke, bai lagun baten edo lagun-talde baten eskutik; baina, era berean, corpus bat automatikoki marka daiteke, konputagailu-programa baten eskutik, eta horren irteera pertsona batek zuzendu dezake edo ez.
5. Azken erabiltzaileak kontuan izan beharko du corpusa ez dela inongo akatsik ez duen zerbait, baizik eta erabilgarri izan ahalko den tresna. Markaketa-prozesu bat, definizioz, interpretazio-prozesu bat izango da, bai testuaren egiturarena, bai testuaren mamiarena.
6. Markatze-eskemak, ahal den heinean, modu zabal batean onartutakoak izan behar dira, eta neutralak teoria mailan. Adibidez, sintagma mailan markatutako corpusak usu testuingururik gabeko

gramatiketan oinarritzen dira, Chomsky-ren *Printzipio eta Parametroak* bezalako egitura hertsiago baten gainean baino areago.

7. Ez da lehenetsiko inongo markaketa-egiturarik estandarra izateko. Estandartasuna praktikotasunak emango dio.

8.5.3.1 Testu markatuak vs testu etiketatuak

Zenbait markaketa linguistiko, hala nola hitzei kode bereziak atxikitzea haien zenbait ezaugarri adierazteko, *markaketa* gisa baino gehiago *etiketatze* (*tagging*) gisa ezagutzen dira; eta ezaugarriei egokitzen zaizkien kodeei *etiketa* (*tag*) esaten zaie. Etiketatzea zenbait kontu markatzeko erabiltzen da. Eta horregatik maila desberdinetako etiketatzeak daudela diogu. Maila horiek, laburki aipatuta, hurrengoak dira:

- **Kategoria mailako etiketatzea.** Corpus linguistikoetan oinarri-oinarrizko etiketatzea da. Helburua da testuko unitate lexikal bakoitzari bere kategoria adieraziko dion kode bat lotzea. Kategoriaren marka oso erabilgarria da corpusetatik berreskura daitekeen informazioaren zehaztasuna handitu egiten duelako, baita ondorengo analisietarako oinarri izango delako ere (hala nola, analisi sintaktikorako eta eremu semantikoaren kodeketarako). Kategoriaren markak kasu homografoak bereizteko lagungarri dira. Lematizaziorako ere oso garrantzitsua da. Lematizazioa kategoria ezagutzearekin oso lotuta dago, eta corpus bateko hitz guztiak haien lema edo erroetara bihurtzea dakar. Lematizazioak ahalbidetzen du ikertzaile batek lexema batek izan ditzakeen aldaketa guztiak ezagutzea, eskuz sartu gabe, eta gainera, lexema horren frekuentzia eta distribuzioa ere ezagutu ahal izango ditu. Analizatzaile morfologikoa egiteko ere behar-beharrezkoa da. Euskara bezalako flexio handiko hizkuntzatan lematizazioa eta egiaztapen ortografikoa hitz-zerrendetan oinarrituta egitea ia ezinezkoa da, lema batetik sor daitekeen forma kopurua izugarri handia baita. Hona, esaterako, IXA taldean euskararako erabiltzen dugun **kategoria-sistema:**

  Kategoria lexikalak (15)

  Kategoria nagusiak (10)

  - IZE IZENAK
    - ARR ARRUNTAK *(zuhaitz)*
    - IZB PERTSONA-IZEN BEREZIAK (Mikel)
    - LIB LEKU-IZEN BEREZIAK (Donostia)
    - ZKI ZENBAKIA (bat)
  - ADJ ADJEKTIBOAK
    - ARR ARRUNTAK (handi, benetako)
    - GAL GALDETZAILEAK (nongo)
  - ADI ADITZAK
    - SIN SINPLEAK (ekarri)
    - ADK KONPOSATUAK (lo egin)
    - ADP PERIFRASTIKOAK (ahal izan)
    - FAK FAKTITIBOAK (etorrarazi)
  - ADB ADBERBIOAK

ARR ARRUNTAK (gaur, negarrez)
GAL GALDETZAILEAK (noiz)

- DET DETERMINATZAILEAK
  - ERK ERAKUSLEAK
  - ERKARR ARRUNTAK (hau)
  - ERKIND INDARTUAK (berori)
  - NOL NOLAKOTZAILEAK
  - NOLARR ARRUNTAK (edozein)
  - NOLGAL GALDETZAILEAK (zein)
  - ZNB ZENBATZAILEAK
  - DZH ZEHAZTUAK (bi)
  - BAN BANATZAILEAK (bina)
  - ORD ORDINALAK (bigarren)
  - DZG ZEHAZTUGABEAK (zenbait)
  - ORO OROKORRAK (guzti)
- IOR IZENORDAINAK
  - PER PERTSONALAK
  - PERARR ARRUNTAK (ni)
  - PERIND INDARTUAK (neu)
  - IZG ZEHAZTUGABEAK
  - IZGMGB MUGAGABEAK (norbait)
  - IZGGAL GALDETZAILEAK (nor)
  - BIH BIHURKARIAK (-(r)en burua)
  - ELK ELKARKARIAK (elkar)
- LOT LOTURAZKOAK
  - LOK LOKAILUAK (hala ere)
  - JNT JUNTAGAILUAK (edo)
- PRT PARTIKULAK (*omen*, *ote*...)
- ITJ INTERJEKZIOAK (*alajaina!*)
- BST BESTELAKOAK (*baldin*)

Kategoria lagungarriak (5)

- ADL ADITZ LAGUNTZAILEAK (*du*)
- ADT ADITZ SINTETIKOAK (*dator*)
- SIG SIGLAK (*EHU*)
- SNB SINBOLOAK (*km, cm, g...*)
- LAB LABURDURAK (*etab.*)

Kategoria morfologikoak (9)

- AMM ADITZ-MOTA MORFEMAK *(-tu, -t(z)e...)*
- ASP ASPEKTU-MORFEMAK (*Ø, -ko...*)
- ATZ ATZIZKIAK (*-pe*)
- AUR AURRIZKIAK (*ber-*)
- DEK DEKLINABIDE-MORFEMAK (*-aren*)
- ELI ELIPSIA (*Ø*)
- ERL ERLAZIO-ATZIZKIAK (*-(e)la*)

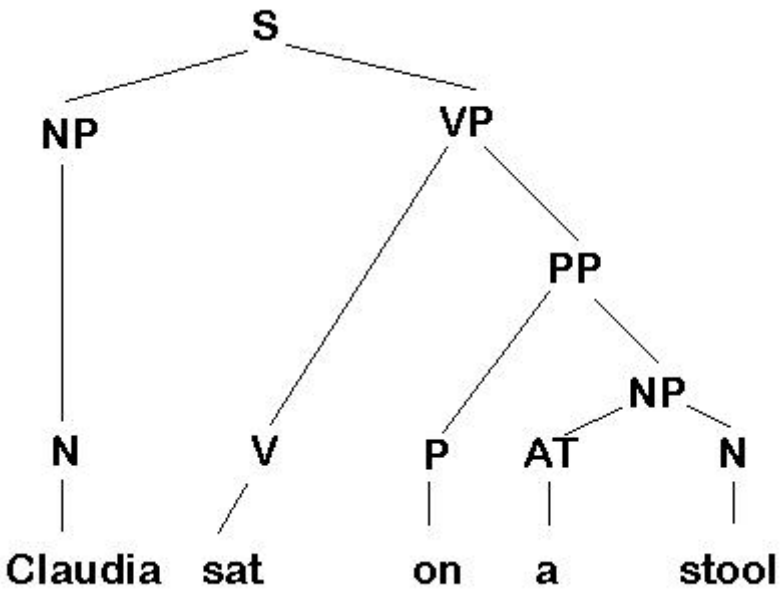
- GRA GRADUATZAILEAK (*-ago*)
- MAR MARRA (-)

Puntuazio-zeinuak (3)

- PNT PUNTUA
- BPM BESTE PUNTUAZIO-ZEINU BATZUK
  (puntuaren pareko izan daitezkeenak)
- PSB PUNTUAZIO-SINBOLOAK
  (parentesiak, marra luzea, kakotxak...)

- **Etiketatze sintaktikoa.** Analizatzaile sintaktikoak oinarrizko kategoria morfosintaktikoak elkarren arteko erlazio sintagmatiko bihurtzen ditu. Hau da, ziurrenik, corpus bat etiketatua ikusteko garaian, kategoria etiketatzailearen ondoren gehien agertzen dena. Sintaktikoki analizatutako corpusak zuhaitz-banku (*treebank*) gisa ezagutzen dira. Hitz horrek adierazten du analizatzaileak erabiltzen dituen sintagma-markatzaileek zuhaitz itxura dutela. Adibidez, "Claudia sat on a stool" (BNC) esaldiak honoko errepresentazioa du:

  (S=sentence, NP=noun phrase, VP=verb phrase, PP=prepositional phrase, N=noun, V=verb, AT=article, P=preposition.)



Honelako diagrama bisualak oso zailak dira corpusen markaketetan ikustea, eta maizago ikusiko ditugu etiketadun parentesi karratuekin emandakoak. Beraz, adibidez, goian sintaktikoki analizatutako esaldia modu honetan agertuko da:

[S[NP Claudia_NP1 NP][VP sat_VVD [PP on_II [NP a_AT1 stool_NN1 NP] PP] VP] S]

Informazio morfosintaktikoa azpimarraren bitartez (_) lotzen zaie hitzei, kategoria lotzen zitzaien modu berean, eta osagaiak sintagmaren hasieran eta bukaeran parentesi karratuak ireki eta ixtearen bitartez markatzen dira: adib.: [S …… S].

Zenbaitetan parentesi karratuetan oinarritutako informazio hau zabaldu egiten da eta zuhaitz-diagramaren itxura gordetzen du (sistema hori *Penn Treebank project*ek erabiltzen du). Adibidez:

```
[S
  [NP Claudia NP]
   [VP sat
       [PP on
           [NP a stool NP]
       PP]
    VP]
S]
```

- **Etiketatze semantikoa.** Bi markaketa semantiko topa ditzakegu:
  - Testuan dauden elementuen arteko erlazio semantikoa markatzea, adibidez, ekintza jakin batzuetan egon daitezkeen egilea edo jasalea. Hau oraintsu hasi da lantzen.
  - Hitzen ezaugarri semantikoak testuan markatzea, gehienbat hitzaren adiera bat edo beste den zehaztea. Honek historia handixeagoa du, 1960 inguruan hasi baitzen.

  Ez dago adostasun unibertsalik markatu beharko liratekeen ezaugarri semantikoen inguruan –izan ere, aurretik egindako markaketa gehienak giza teoria zientifikoak eraginda izan ziren, adibidez, giza elkarrekintza–. Hala ere, Sedelow-k eta Sedeow-k (1969) *Roget's Thesaurus*-a erabili zuten –bertan hitzak kategoria semantiko orokorretan sailkatzen dira–.

  Hurrengo adibidearekin markaketa motaren adibidea ikusiko dugu, eta zein kategoria semantiko erabiltzen diren:

and	00000000
soldiers	23241000
platted	21072000
a	00000000
crown	21110400
of	00000000
thorns	13010000
and	00000000
put	21072000
it	00000000
on	00000000
his	00000000
head	21030000
and	00000000
they	00000000
put	21072000
on	00000000

him	00000000
a	00000000
purple	31241100
robe	21110321

Kode numerikoek hau adierazten dute:

00000000	Low content word (and, the, a, of, on, his, they etc)
13010000	Plant life in general
21030000	Body and body parts
21072000	Object-oriented physical activity (e.g. put)
21110321	Men's clothing: outer clothing
21110400	Headgear
23231000	War and conflict: general
31241100	Colour

Kategoria semantikoak 8 zenbakiz adierazi ohi dira. Goiko hau Schmidt-ek (1993) erabilitakoa da, eta egitura hierarkikoa du. 3 *top level*en gainean eraikita dago, eta era berean, azpimultzoak dituzte.

- **Diskurtsoaren eta testuaren etiketa linguistikoak.** Corpusetan gutxien erabiltzen diren markaketa motak testu eta diskurtso mailakoak dira. Hala ere, zenbaitetan marka hauek topa ditzakegu: *Diskurtso* mailakoak eta *anafora*ri dagozkionak.

  Diskurtso mailan, Stenström-ek (1984) *London-Lund* ahozko corpusa 16 diskurtso-mailako etiketekin markatu zuen. Modu honetako kategoriak txertatu zituen:

  "apologies" e.g. sorry, excuse me

  "greetings" e.g. hello

  "hedges" e.g. kind of, sort of thing

  "politeness" e.g. please

  "responses" e.g. really, that's right

  Nahiz eta potentzialki diskurtsoaren analisian etiketa-mota honek indar handia izan, ez da gehiegi erabiltzen, seguru asko kategoria linguistiko hauek testuinguruarekin oso lotuta daudelako, eta horiek identifikatzea beste fenomeno linguistikoak baino zailagoa izan daitekeelako.

  Markaketa anaforikoari dagokionez, kohesioari esker testuan ditugun hitzek elkarren arteko lotura agertzen dute, eta hori izenordainen, errepikapenen, aldaketen eta horrelakoen bitartez lortzen da. Halliday eta Hasan-en *Cohesion in English* (1976) linguistikan oso kontuan hartutakoa da, kohesioari buruzko gai garrantzitsuak aztertzen dituelako.

  Izenordainek aurreko elementuei egiten diete erreferentzia (anafora), eta hori ezagutzeko datu enpiriko asko eta asko beharko lirateke… beste hitzetan esanda, corpus handiak.

Marka anaforikoak, oraingoz, pertsonek bakarrik jar ditzakete, nahiz eta helburuetako bat informazio honekin makinak entrenatzea den, ondoren makinek egin dezaten lan. Corpus-zatitxo txiki batzuk besterik ez daude modu honetan markatutakoak: hauetako bat *Lancaster/IBM treebank* anaforikoa dugu, eta hona adibidea:

> A039 1 v (1 [N Local_JJ atheists_NN2 N] 1) [V want_VV0 (2 [N the_AT (9 Charlotte_N1 9) Police_NN2 Department_NNJ N] 2) [Ti to_TO get_VV0 rid_VVN of_IO [N 3 <REF=2 its_APP$ chaplain 3) ,_, [N {{3 the_AT Rev._NNSB1 Dennis_NP1 Whitaker_NP1 3} ,_, 38_MC N]N]Ti]V] ._.

Goiko testuak kategoria mailako etiketak ditu, azaleko sintaxiari dagozkionak eta etiketa anaforikoak. Kodeen azalpena honokoa da:

> (1 1) etc. - noun phrase which enters into a relationship with anaphoric elements in the text
>
> <REF=2 - referential anaphor; the number indicates the noun phrase which it refers to – here it refers to noun phrase number 2, the Charlotte Police Department
>
> {{3 3}} - noun phrase entering in equivalence relationship with preceding noun phrase; here the Rev Dennis Whitaker is identified as being the same referent as noun phrase number 3, its chaplain

- **Etiketatze edo transkripzio fonetikoa.** Ahozko corpusak transkripzio fonetikoa erabiliz kodetu daitezke. Oraingoz ez ditugu modu honetan transkribatutako corpus publiko asko. Arrazoia izan daiteke oraingoz lan hau konputagailuek ezin dutela egin, eta pertsonek egin behar izaten dutela. Pertsona horiek, gainera, oso kualifikatuak izan behar dira soinuak jaso eta transkribatzeko garaian. Beraz, denbora dezente eraman dezakeen kontua dugu hau. Beste arazo bat da hitzen transkripzio fonetikoa egitean hitzetan gelditzen garela, eta egia esan, soinu hauen muga ez da hain zehatza izaten; izan ere, zenbaitetan soinu bera izan beharko luketen bi hizkik soinu ezberdina izan dezakete testuingurua ezberdina delako.
  Hala ere, fonetikoki transkribatutako corpusak oso erabilgarriak dira azterketa fonetikoak egiteko laborategietako tresnak ez dituztenentzat. Adibide bat *MARSEC* corpusa dugu (*Lancaster/IBM* ahozko corpusa du oinarri) eta Lancaster eta Leeds-eko unibertsitateek landu dute. *MARSEC* corpusak transkripzio fonetikoa du.
- **Prosodiari loturiko etiketatzea.** Prosodiak soinuari egiten dio erreferentzia, baina ez hitz mailakoari, baizik eta maila goragokoari, hala nola, indarra, intonazioa eta erritmoa. Prosodia mailan markatutako corpus normalean era zabalean onartutako moldeak erabiltzen dituzte. Halakoa da O'Connor eta Arnold-ena (1961). Oro har, markatzen diren intonazioak nabarienak dira, silaba bakoitzaren intonazioa baino gehiago. Honoko adibidea *London-Lund* corpusetik hartuta dago.

  1 8 14 1470 1 1 A 11 ^what a_bout a cigar\ette# . /
  1 8 15 1480 1 1 A 20 *((4 sylls))* /
  1 8 14 1490 1 1 B 11 *I ^w\on't have one th/anks#* - - - /

1 8 14 1500 1 1 A 11 ^aren't you .going to sit d/own# - /

1 8 14 1510 1 1 B 11 ^[/\m]# - /

1 8 14 1520 1 1 A 11 ^have my _coffee in p=eace# - - - /

1 8 14 1530 1 1 B 11 ^quite a nice .room to !s\it in ((actually))# /

1 8 14 1540 1 1 B 11 *^\isn't* it# / 1 5 15 1550 1 1 A 11 *^y/\es#* - - -

Adibide honetan erabilitako kodean honokoak dira:

end of tone group

^ onset

/ rising nuclear tone

\ falling nuclear tone

/\ rise-fall nuclear tone

_ level nuclear tone

[] enclose partial words and phonetic symbols

. normal stress

! booster: higher pitch than preceding prominent syllable

= booster: continuance

(( )) unclear

* * simultaneous speech

- pause of one stress unit

Corpus bat prosodiari begira etiketatzeak, hala ere, hainbat arazo ditu:

- Juizioak normalean inpresio baten ondorio dira. Adibidez, tonu baten mailaren mugimendua adosteko zailtasuna duen gaia da. Zenbaitek doinuaren beherakada sumatuko duten leku berean, beste batzuek beherakada horren ondoan halako gorakada suma dezakete… honek bigarren puntura garamatza.
- Zaila da konsistentzia izatea, gehienbat corpusa pertsona batek baino gehiagok lantzen badute. Hau arindu ahal izango da bi lagun jarriko bagenitu corpusaren zati txikitxo berdina etiketatzen.
- Testua berreskuratzea zaila da. Leech-en lehen maximak zioenez, testu baten markak kenduz gero, erraza izan behar da testu gordinera itzultzea, eta hori ez da horrela maila prosodikoan markatutako testu batean. Prosodia-markak hitz baten barruan agertu ahal izango dira, eta horrek zailtzen du testu gordinera itzultzea.

Zenbaitetan prosodia-kontu batzuk markatzeko karaktere grafiko bereziak erabiltzen dira. Eta konputagailu eta inpresora guztiek ezin dituzte karaktere horiek ezagutu. Beraz, TEI gidalerroek arazoa arindu egingo dute hein batean.

- **Arazoei zuzendutako etiketatzea.** Arazo bati zuzendutako etiketatzea (Haan-ek (1984) agertu zuen bezala), erabiltzaileak corpus bat hartzea (etiketatuta dagoena edo ez) eta bere helbururako beharrezko dituen ezaugarriak jartzea litzateke. Orain arte ikusitako etiketatze motekin bi diferentzia nagusi ditu:

- Ez da erabat zehatza. Hitz edo perpaus guztiak ez dira etiketatuko, baizik eta ikerketarako garrantzitsuak diren horiek soilik. Hau da arazoei zuzendutako etiketatzea eta anaforaren etiketatzea batzen dituen zerbait.
- Etiketatze-eskemak aukeratu egiten dira, ez batez bestekoei edo teoriaren neutraltasunari begira, baizik eta ondoren egingo den ikerketarako behar diren galderei erantzuteko interesgarri izango diren ezaugarriei begira.

Nahiz eta etiketatze mota honi buruz zaila den gehiago orokortzea, kontuan izatekoa da, zenbaitetan lan oso praktikoak egin baitaitezke corpus zabalak erabiliz etiketatze mota honekin.

8.5.3.2 Corpusak markatzeko estandarrak

Gaur egun testuetako informazioa irudikatzeko ez dago era zabal batean onarturik dagoen estandarrik, eta atzera begira hainbat hurbilpen erabili izan dira, batzuk besteak baino iraunkorragoak izan direnak. Aspaldiko markaketa-modua *COCOA* da. *COCOA* hasierako konputagailu-programa zen eta MRD testuetatik markatutako hitzak eta testuingurua ateratzeko erabiltzen zen. Honek erabiltzen zituen gomendioak beste programa batzuetan erabili ziren gerora, hala nola, *OCP* (*Oxford Concordance Program*). *Longman-Lancaster* corpusek eta *Helsinki* corpusek ere *COCOA*-ren gomendioak erabili dituzte.

Oso modu sinplean, *COCOA* markak bi erreferentzia izango ditu angeludun parentesien barruan:

Aldagai baten izen jakina ordezkatzen duen kodea.

Aldagai horrek adierazten dituenen loturak edo lotura-multzoak.

Adibidez, A kodea “egile” aldagaia adierazteko erabil daiteke, eta lotura egilearen izenarekin izango du.

<A CHARLES DICKENS>

<A WOLFGAN VON GOETHE>

<A HOMER>

COCOA loturak testuen informazioaren tipo zehatzak kodetzeko joera informal bat besterik ez du adierazten, hala nola, egileak, datak eta tituluak. Gaur egungo joerek gehiago bilatzen dute nazioarteko kodeketa estandar formalizatuagoa. Egungo joera honen erregea *Text Encoding Initiative* (TEI) dugu; joera hau Association for Computational Linguistics, the Association for Literary and Linguistic Computing eta the Association for Computers and the Humanities elkarteek bultzatu dute. Helburua, inplementazio estandarrak lortzea da, makinak irakur ditzakeen testuen arteko trukatzea ahalbidetzeko.

TEIk erabiltzen duen markatze-sistema SGML da (*Standard Generalised Markup Language*). SGMLk honoko abantailak ditu:

Argitasuna

Sinpletasuna

Forma zehatza

Nazioarteko estandar gisa onartua izatea

TEIren ekarpena da arau zehatz batzuk ezartzea, testuak markatzeko estandar hau nola erabili erakusteko (Sperberg-McQueen eta Burnard, 1994).

TEIn testu (edo dokumentu) bakoitzak bi atal ditu, burukoa eta testua bera. Burukoak honelako informazioa du:

Egilea, titulua eta data

MRDa sortzeko erabilitako bertsioa

Erabilitako markaketa-moduari buruzko informazioa

8.5.3.3 Corpus eleaniztunak

Corpus eleaniztunak hizkuntza batzuetako corpusen bildumak dira. Bi mota bereizten dira:

- Bata, corpus elebakarretatik ateratako testu multzoen bilduma da. Bilduma honetan hizkuntza bakoitzeko zenbait testu ditugu; hots, ez dio axola testu beraren itzulpenak ez izateak. Adibidez, danieraren, frantsesaren eta ingelesaren *Aarthus* corpusak hiru hizkuntzetako arauak biltzen ditu, baina ez dira testu beraren itzulpenak. Testu guztiak ezaugarri berekoak direnean (gaia, urtea, mota...), *corpus konparagarriak* direla esaten dugu.
- Bestea, garrantzi handiagoa duena, *corpus paraleloak* dira, bitestuak ere deituak. Corpus hauetan testu bera hizkuntza bat baino gehiagotan ematen da. Corpus paralelo hauek ez dira gaur goizekoak, Erdi Aroan ere bai baitzituzten hainbat hizkuntzatan emandako Bibliak (hebreeraz, latinez eta grekoz). Baina corpus paralelo bat ez da berehala erabil daitekeen corpusa. Corpus hauek erabilgarri izan daitezen, jakin behar da hizkuntzen arteko unitate linguistikoak nola lotzen diren elkarri. Hau eginda duen corpus motari *corpus parekatua* esaten zaio (noizbait *corpus lerrokatua* ere deitua). Esaldiz esaldi parekatutako corpusetan bata bestearen itzulpena diren esaldiak markatzen dira. Perpausaz gain badira beste parekatze maila posibleak: hitzak, terminoak, lokuzioak, kolokazioak, entitateak. Adibidez, corpus batean "*Das Buch ist auf dem Tisch*" eta "*The book in on the table*" perpausek lerrokatuta egon beharko dute, adibidez, "*Das*" "*The*"rekin. Hau ez da beti prozesu sinplea; izan ere, hizkuntza bateko hitz bat beste hizkuntza bateko birekin lotu beharko da, adibidez, alemaneko "*raucht*" ingeleseko "*is smoking*" hitzekin lotu beharko da. Eta aurrekoari euskarazko adibidea jarriko bagenio ("liburua mahai gainean dago"), ingelesez izenaren aurretik doan artikulua ("the") euskaraz izenaren atzetik eta horri lotuta doan atzizkiarekin ("-a") lotu beharko litzateke. Bi hizkuntza horien artean preposizio eta postposizioen arteko gurutzaketak gertatzen dira.

Gero eta lan handiagoa egiten ari da corpus elebidunak eraikitzeko, eta batez ere, corpus paraleloak eraikitzeko, hauek baitira Lengoaia Naturalaren Prozesamendurako baliagarrienak. Gaur egun, corpus paraleloen bi adibide daude, eta hizkuntza anitzekoak izan beharrean, bi hizkuntza besterik ez dituzte. Hala eta guztiz ere, Europar Batasunak lagundutako bi proiektu (CRATER eta MULTEXT) saiatzen ari dira hizkuntza anitzeko testu paraleloak sortzen. *Canadian Hansard* corpusa etiketatuta dago, eta baditu

testu paraleloak ingelesez eta frantsesez, baina testu-mota batzuk bakarrik biltzen ditu (kanadiar parlamentuko *proceeding*ak). Hala ere, bide hau handitzen ari den mundua da, eta baliteke egoera goitik behera aldatzea urte gutxitan.

Hona corpus elebidun baten adibidea (adibide hau frantses-ingeles corpus paralelo batetik hartutako adibidea da eta perpaus mailako loturak ditu):

sub d = 22 ----------&

the location register should as a minimum contain the following information about a mobile station : -----&

1 ' enrigisteur de localisation doit contenir au moins les renseignments suivants sur une station mobile : sub d = 386 ----------&

handover is the action of switching a call in progress from one cell to another (or radio channels in the same cell) . -----&

le transfert intercellulaire consiste à commuter une communication en cours d ' une cellule (ou d ' une voie radioélectrique à l ' autre à l ' intérieur de la même cellule) . sub d = 380 ----------&

the location register, other than the home location register used by an msc to retrieve information for, for instance, handling of calls to or from a roaming mobile station , currently located in its area . -----&

enregistreur de localisation , autre que l ' enregistreur de localisation nominal , utlilisé par un ccm pour la recherche d ' informations en vue , par exemple , de l ' établissment de communication en provenance ou à destination d ' une station mobile en déplacement , temporairement située dans sa zone .

9 Hizkuntza-produktuak garatzeko estrategia

Hizkuntza-teknologiak funtsezkoak dira Informazio eta Komunikazioaren Gizartea esaten dugun honetan. Eusko Jaurlaritzako hiru sailek batera, Industriak, Hezkuntzak eta Kulturak, ikerlerro estrategiko gisa definitu izan dute 2002. urtetik aurrera. Testuinguru horretan kokatzen den Etortek 2002-04 ikerketa-deialdian, VICOMTech, Elhuyar, Robotiker, Aholab eta IXA taldeak proiektu batean dihardugu: "*HIZKING21 HIZKuntza INGeniaritza XXI. mendeko atarian*" (http://www.hizking21.org). Ingeniaritza linguistikoan, ikerketan, eta garapenean lan egiteko epe erdirako estrategia diseinatu dugu, proiektuko partaideen 15 urteko eskarmentuan oinarrituta. Nazioartean, puntako mailan mugituko den industria sendoa sortzea da gure erronka. Ikerketa-taldeak, industriak eta erakunde ofizialak koordinatu egin behar dira helburu hori lortzeko. Ingeniaritza linguistikoan, Ikerketan eta Garapenean arituko den komunitate zabal bat sortu behar dugu. 2003. urtean, estimatzen genuen 120-150 pertsona zebilela lanean Euskal Herrian hizkuntzaren prozesamenduaren gainean, zuzenki edo zeharka. Gauzak ondo planifikatuz gero eta formazio-plan egokiak aurrera eramanez gero, urte gutxi barru kopuru hori bikoiztu edo hirukoiztu egin daitekeela aurreikus daiteke. Baina, nola ekin erronka horri? IXA taldeak urteetan garatu izan du estrategia bat, urrats-kate bat, hizkuntzaren teknologiari metodo batekin ekiteko. Taldearen 15 urteko ibilbidea estrategia horren arabera egin da. Nazioarteko foroetan ere aurkeztu eta kontrastatu izan da beste ikertzaile batzuekin. Ideia nagusiak oso sinpleak dira :

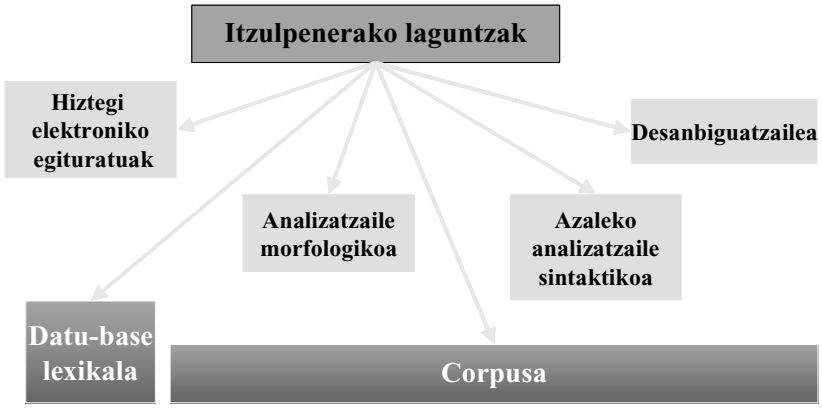
- Hasieran sortu behar dira oinarrizko baliabideak eta tresnak, eta geroago sortu merkatu-aplikazioak. Ez da egin behar alderantziz!
- Oinarri linguistiko bakoitza, tresna eta aplikazio bakoitza ondo diseinatu behar da ondorengo produktuetan erabilgarria izan dadin.

Produktuen atala aurkezterakoan esan den modura, aplikazioen garapenerako oinarri sendo batetik abiatu beharra dago. Oro har, hizkuntza-teknologien egitura, piramide moduko batez irudika daiteke (ikus 12. irudia). Piramide horren oinarrian ingeniaritza linguistikoan lan egiteko beharko diren oinarrizko baliabideak daude. Baliabide horiei esker, tresnak garatzeko modua izango da, eta behin tresnak garatuta, ingeniaritza linguistikoaren hainbat arlotan lan egiteko moduko produktu komertzialak kaleratu ahal izango ditugu. Kontuan izan behar da, ordea, alderantzizko bidea ezin dela egin, etxea teilatutik eraiki nahi ez badugu.

Aplikazioak garatuko badira, zer-nolako azpiegitura behar da?

Aplikazioak ditugu helburu, noski. Gizarte eleaniztun batean bizi gara, eta eleaniztasun horretan lagungarri izango zaizkigun tresnekin egiten dugu amets: euskararako itzulpen automatikoa, hizketaren

ezagutza, estilo-zuzentzaileak. Baina horiek sortzera helduko bagara, oinarri sendo bat behar dugu lehenik. Esaterako, itzultzaileentzat lagungarri izan daitekeen tresna baten garapenerako hainbat baliabide eta tresna garatu beharko ditugu lehenik (ikus 46. irudia), baina baliabide eta tresna horiek, beste guztiak bezala, itzulpena ez diren beste aplikazioetan ere erabilgarri izan beharko dira (ikus 47. irudia).

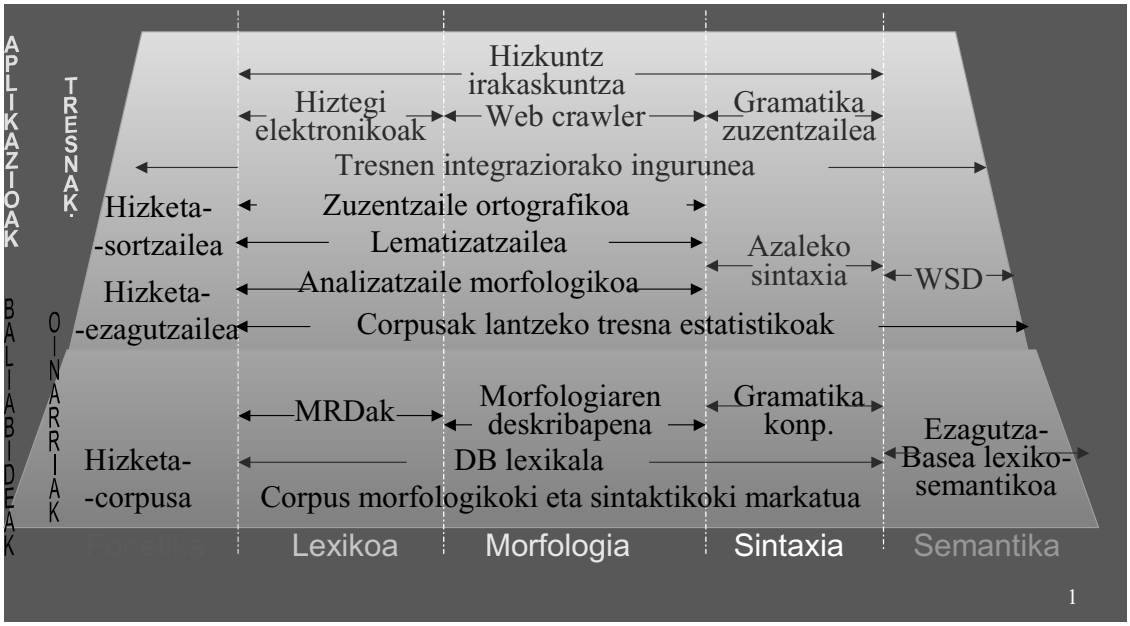


46. irudia. Aurretik sortutako baliabide eta tresnen berrerabilera itzulpenerako laguntzak egiterakoan

Produktu bakoitza, produktu berrien garapenean ahalik eta modu zabalenean berrerabiltzea da helburua. Horrela, gaur eguneko lorpenak eta jarduerak 47. irudian agertzen dira.

Badira hainbat produktu euskara eta softwarea uztartzen dituztenak. Euskararen Software Katalogoan (www.ueu.org/softkat) 140 bildu dira. Horietarik 34 Hizkuntzaren Industriarekin lotuta daude. Hori ez da hutsaren hurrengoa, baina bai oso gutxi; ahalegin handia egin behar dugu informazioaren gizarteko mundu honetan euskara atzean ez gelditzeko.

Bide horretan sortuko den oinarri linguistiko bakoitza, tresna eta aplikazio bakoitza, ondo diseinatu beharko da ondorengo produktuetan erabilgarria izan dadin. Hori da erronka.



47. irudia. Egun euskararako dauden zenbait baliabide linguistiko, tresna eta aplikazio

10 Bibliografia

Aarts J. & Meijs W. (arg.) 1986. *Corpus Linguistics II*, Amsterdam: Rodopi.

Abaitua J., Aduriz I., Agirre E., Alegria I., Arregi X., Artola X., Arriola J.M., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar M., Sarasola K., Urkia M., Zubizarreta J.R. 1992. *Estudio comparativo de diferentes formalismos sintacticos para su aplicacion al euskara*. Barne-txostena, UPV/EHU/LSI.

Abaitua, J. 1988. *Complex predicates in Basque: from lexical forms to functional structures*. Doktore-tesia, University of Manchester.

Aduriz I., Arriola J.M. & Díaz de Ilarraza A. (2003) Desanbiguazio morfologikoa, azterketa sintaktikoaren lehen urratsak eta aplikazioak Murriztapen Gramatikaren eredu konputazionala jarraituz. In J.M. Makazaga & B. Oyharçabal (arg.) *Euskal gramatikari eta literaturari buruzko ikerketak XXI. mendearen atarian*. Gramatika gaiak, Iker-14. Euskaltzaindia, Baiona.

Aduriz I. & Ilarraza A. (2003) Morphosyntactic disambiguation and shallow parsing in computational processing of Basque. In B. Oyharçabal (arg.) *Inquiries into the lexicon-syntax relations in Basque.* ASJUren gehigarria. Euskal Herriko Unibertsitatea. Bilbo.

Aduriz I. 2000. *EUSMG: Morfologiatik sintaxira Murriztapen Gramatika erabiliz*. Doktore-tesia. Euskal Filologia Saila. Euskal Herriko Unibertsitatea.

Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 1999. *MORFEUS: Euskararako analizatzaile morfosintaktikoa*. Barne-txostena, UPV/EHU/LSI/TR 1-99.

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. 1997. A spelling corrector for Basque based on morphology. Literary & Linguistic Computing, Vol. 12, No. 1. Oxford University Press. Oxford. 1997.

Agirre E., Ansa O., Arregi X., Artola X., Díaz de Ilarraza A., Lersundi M., Martínez D., Urizar R., Sarasola K.. Extraction of semantic relations from a Basque monolingual dictionary using Constraint Grammar, in *Proceedings of Second Int. Conf. on Language Resources and Evaluation*. Atenas, 2000.

Agirre E. 1999. *Kontzeptuen arteko erlazio-izaeraren formalizazioa ontologiak erabiliaz: Dentsitate Kontzeptuala.* Doktore-tesia. Lengoaia eta Sistema Informatikoak Saila. Euskal Herriko Unibertsitatea.

Agirre E., Agirre A., Alegria I., Arregi X., Artola X., Diaz de Illarraza A., Goenaga P., Maritxalar M., Sarasola K., Urkia M. Bi mailatako morfologiaren euskararako egokitzapena, Elhuyar, 17. lib., 6-14. 1991.

Agirre E., X. Arregi, X. Artola, A. Díaz de Ilarraza, F. Evrard, K. Sarasola (1994b). A methodolgy for the extraction of semantic knowledge from dictionaries using phrasal patterns. *Proc. of IBERAMIA'94*, 263-270. Caracas (Venezuela).

Agirretxe J.L, Lersundi M., Olaetxea O. 1998. *Pasaiako hizkera*. Pasaiako Udala. Euskara Batzordea.

Ait-Mokhtar, Salah & Jean-Pierre Chanod. 1997. Incremental finite-state parsing. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 72-79.

Aijmer K. and Altenberg B. (arg.) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Londres: Longman.

Aldezabal I. 2004. *Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz.* Euskal filologia saila. Zientzia fakultatea. Leioa. UPV/EHU. 2004ko apirila.

Aldezabal I., Gojenola K., Sarasola K. 2003. *Baterakuntzan oinarritutako euskararen analizatzailea. Oinarrizko PATR gramatika*. In J.M. Makazaga & B. Oyharçabal (arg.) *Euskal gramatikari eta literaturari buruzko ikerketak XXI. mendearen atarian*. Gramatika gaiak, Iker-14. Euskaltzaindia, Baiona.

Aldezabal I., Ansa O., Artola X., Ezeiza A., Gojenola K., Insausti J.M., Lersundi M. 1999. *Euskararen Datu-Base Lexikala: eskema berriaren proposamena.* UPV/EHU / LSI / TR 9-99 barne-txostena. Donostia.

Alegria I. 1995. *Euskal morfologiaren tratamendu automatikorako tresnak.* Doktore-tesia. UPV-EHU.

Alegria I., Artola X., Sarasola K. 1997. Hizkuntzaren tratamendu automatikoa: helburuak eta abiaburuak. JAKIN 102 zk., 61-82.

Alegria I. & Urkia M. 2002. *Morfologia konputazionala. Euskararen morfologiaren deskribapena*. Udako Euskal Unibertsitatea. Bilbo.

Allen J., Hunnicutt M., Klatt D. 1987. *From text to speech: the MITalk System*. Cambridge University Press.

Allen J. 1988. *Natural Language Understanding.* The Benjamin/Cummings Publish Company, California.

Allen J. 1995. *Natural Language Understanding. Second edition.* Benjamin Cummings Publishing Company.

Allen, J., Hunnicut, S., Klatt, D. 1987. From Text To Speech, The MITALK System. Cambridge University Press, Cambridge.

Alshawi H. & Crouch R.. 1992. Monotonic semantic interpretation. *Proc. 30th ACL. 1*992

Amsler R.A. 1989. Research Toward the Development of a Lexical Knowledge Base for Natural Language Processing. SIGIR 1989: 242-249.

Antworth E.L. 1990. *PC-KIMMO: a two-level processor for morphological analysis*. Occasional Publications in Academic Computing, No. 16, Dallas, Texas.

Antworth E. L. 1991. "Introduction to two-level phonology." Notes on Linguistics 53: 4-18.

Arregi X. 1995 *Anhitz: Itzulpenean laguntzeko hiztegi-sistema eleanitza*. Doktore-tesia. Euskal Herriko Unibertsitatea

Arriola J. M. 2000. *Euskal Hiztegiaren azterketa eta egituratzea ezagutza lexikalaren eskuratze automatikoari begira. Aditz-adibideen analisia Murriztapen-gramatika baliatuz, azpikategorizazioaren bidean*. Doktore-tesia. Euskal Herriko Unibertsitatea: Gasteiz.

Artola X. 1993. *Hiztegi-ezagumenduaren errepresentazioa eta arrazonamenduaren ezarpena.* Doktore-tesia. Informatika Fakultatea, UPV-EHU.

Atkins B. T. S. & Levin B. 1995. Building on a corpus: a linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography* 8:2, 85-114.

Atkins B.T.S. & Zampolli A. (arg.). 1994. *Computational Approaches to the Lexicon.* New York: Oxford University Press.

Aurnague M. Cas inessif du basque et connaissance du monde: l'expression de l'espace a-t-elle horreur du vide (sémantique) ?

Aurnague M. *Orientation in French spatial expressions: formal representations and inferences*

Aurnague M. *Euskal "genitiboen" semantikarantz meronomien ildotik*

Atserias J., Castellón I, Civit M.. 1998. “Syntactic parsing of unrestricted spanish text”, *Proceedings of 1st international conference on language resources and evaluation (LREC)*, Granada.

Atwell E.S. How to detect grammatical errors in a text without parsing it. In *ACL Proceedings, Third European Conference* 1987: 34-45.

Axular 1634.. Jakin. 1976. Arantzazu.

Berwick R.C. & Brent M.R. 1991. Automatic acquisition of subcategorization frames from tagged text. *Speech and Natural Language: Proceedings of a Workshop*. Pacific Grove. California.

Barton G., Berwick R., & Ristad E. 1987. *Computational Complexity and Natural Language*. MIT Press.

Beesley K., Karttunen L. 2002. *Finite State Morphology: Xerox Tools and Techniques*. Cambridge University Press.

Bertuccelli, M. 1993. *Che cos'e la pragmatica.* Milán: Bompiani. (Gazteleniazko itzulpena: *Qué es la pragmática.* Bartzelona: Paidós, 1996.)

Bert Esselink. 1998. *A practical guide to software localization*. John Benjamins.

Bessley K & Karttunen L. 2003. Finite State Morphology. *CSLI Publications*. Stanford.

Bessley K. & Karttunen L. 2000. Finite-State for Non-Concatenative Morphotactics. *Proceedings of the ACL'2000*. Hong Kong..

Black W.J. & El-Kateb S. 2004. "A prototype English-Arabic dictionary based on WordNet", *Proceedings of the Second Global WordNet conference (GWC)*: 67-74., Brno, Czech Republic. t

Black A., van de Plassche J., Williams B. 1991. Analysis of Unkown Words through Morphological Descomposition. Proceeding of 5th Conference of the EACL, vol. 1, 101-106.

Bod R.1993.Using an Annotated Language Corpus as a Virtual Stochastic Grammar. *Proceedings of the 11th National Conference on Artificial Intelligence*. Washington, DC, USA. The AAAI Press/The MIT Press, ISBN 0-262-51071-5: 778-783

Bod R. & Kaplan R.M. 1998. A Probabilistic Corpus-Driven Model for Lexical-Functional Analysis. COLING-ACL 1998: 145-151

Boguraev B., Briscoe T. (arg.). 1989. *Computational Lexicography for Natural Language Processing.* New York: Longman.

Boguraev B., Pustejovsky J. (arg.). 1996. *Corpus Processing for Lexical Acquisition.* Massachusetts: The MIT Press.

Breiman L., J. Freidman, R. Olshen & C. Stone. 1984. *Classification and Regression Trees*. Wadsworth, Belmont.

Bresnan J. & Kaplan R.M. 1982. Introduction: Grammars as Mental Representations of Language. Bresnan J., ed., *The Mental Representation of Grammatical Relations*. Cambridge, Massachusetts: MIT Press.

Brill E. 1995. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech TaggingComputational Linguistics, Volume 21: 543-565.

Brill E. & Wu J. 1998. Classifier Combination for Improved Lexical Disambiguation. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*.

Bustos E. 1987. *Filosofía contemporánea del lenguaje I (Semántica Filosófica).* Madril: UNED.

Byrd R.J., Calzolari N. Chodorow M.S, Klavans J.L., Neff M.S. & Rizk O.A. 1987. Tools and methods for computational lexicology. *Computational Linguistics*, v.13 n.3-4.

Byrd R. 1989. "Discovering relationship among word senses", *Proceedings of the 5th annual conference of the UW centre for the New OED*: 67-79. Oxford.

Calzolari N. 1986. "Structure and access in an automated lexicon and related issues", *Workshop automating the lexicon*, Grosseto.

Calzolari N. 1990. (Calzolari, 1990a) "Structure and access in an automated lexicon and related issues", *Computational lexicology and lexicography, Special Issue dedicated to Bernard Quemada*, Pisa. 1. liburukia, 139-161.

Calzolari N. 1990. (Calzolari, 1990b), *Trends in Computational Lexicography and natural language processing*, read at the X reunión anual de la sociedad española del procesamiento del lenguaje natural (SEPLN), Donostia.

Cahill L. 1990. Syllable based morphology. *Proceedings of 13th COLING*, vol. 3, 48-53.

Cahill L. 1993. Morphonology in the lexicon. In: *Sixth Conference of the European Chapter of the Association for Computational Linguistics*. Utrecht. 87-96.

Carballar J.A. 1999. *Internet. Libro del navegante*, RA-MA.

Carreras X., Chao I., Padró L. & Padró M. 2004. FreeLing: An Open-Source Suite of Language Analyzers, in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, PortugaCarrol J. 1993. *Practical unification-based parsing of natural language*. Computer Laboratory. Cambridge University, UK. PhD. thesis. Technical Report 314.

Carrol G. & Rooth M. 1998. Valence Induction with a Head-Lexicalized PCFG. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Granada.

Castellón I. 1992. *Lexicografía Computacional: Adquisición automática de conocimiento léxico*, Universidad de Bartzelona. Doktore-tesia.

Chanod J.P. & Tapanainen P. 1996. (Chanod & Tapanainen, 1996a). A Non-deterministic Tokeniser for Finite-State Parsing. *ECAI'96 workshop on Extended finite state models of language*. Budapest.

Chanod J.P. & Tapanainen P. 1996. (Chanod & Tapanainen, 1996b). A Robust Finite-State Grammar for French. *ESSLLI'96 workshop on Robust Parsing*. Prague.

Charniak E. 1993. *Statistical Language Learning*. Cambridge, M.A: MIT Press.

Chomsky N. 1970. "Remarks on Nominalization.". Reprinted in Davidson and Harman: 262-289.

Chomsky, N. 1997. *Mintzairari buruzko gogoetak,* Donostia, Gaiak.

Church K. 1998. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Procs. of 2nd Conference on Applied Natural Language Processing (ANLP'98)*: 136-143. ACL.

Clark H. H. 1996. *Using language.* Cambridge: Cambridge University Press.

Code & Syntax: Internacionalización, localización y webs multilingües, EITE 2001 (Zentro teknologikoen EITE sarearentzat egindako txosten enkarguzko bat, interesaren arabera, kopiak emango ditugu).

Cole R., Mariani J., Uszkoreit H., Varile G.B, Zaenen A., Zampolli A. 1998. *Survey of the State of the Art in Human Language Technology* Cambridge University Press.

Small S.L., Cottrell G.W. & Tanenhaus M.K. (arg.) 1989. Distributed representations of ambiguous words and their resolution in a connectionist network. *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence.* San Mateo, CA: Morgan Kaufmann Publishers.

Cruse D.A. 1996. *Lexical Semantics.* Londres: Cambridge University Press.

Croft, W. B. 1995. "What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems)", D-lib Magazine.

Dale R., Moisl H., 2000. Somers H. *Handbook of Natural Language Processing*, Marcel Dekker, New York.

Doug Arnold, Balkan, L., Meijer, S., Humphreys, R.L. Sadler, L. 1993. *Machine Translation: An Introductory Guide*. http://clwww.essex.ac.uk/MTbook/

Pereira F., Thishby N., Lee L. 1993. "Distributional clustering of English words". *Proceedings of the 31st Annual meeting of the association for computational linguistics (ACL)*: 183-190. Columbus, Ohio.

Elhuyar (2000) Elhuyar hiztegia. Euskara-gaztelania / castellano-vasco. Elhuyar Kultur Elkartea. Usurbil, Gipuzkoa.

Escandell V. 1993. *Introducción a la pragmática.* Bartzelona: Anthropos.

Esselink B. 1998. A Practical Guide To Software Localization. Amsterdam: John Benjamins Publishing.

Etxepare B. 1545. *Linguae Vasconum Primitiae*.

Euskaltzaindia .2000. *Hiztegi Batua*. Euskaltzaindia, Bilbo.

Euskaltzaindia. 1993. *Euskal Gramatika Laburra: Perpaus Bakuna*. Euskaltzaindia, Bilbo.

Evans R., Gazdar G. 1996. DATR: A Language for Lexical Knowledge Representation, *Computational Linguistics*, vol 22 n. 2, 167-216.

Evans R., Kilgarriff A. 1995. *MRDs, Standards and How To Do Lexical Engineering.* Proceedings, Second Language Engineering Convention, 125-132. Londres.

Ezeiza N. 2002. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua*. Doktore-tesia. Informatika fakultatea, UPV-EHU.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R., 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. COLING-ACL'98, Montreal.

Fellbaum C. 1998. *WordNet, An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Flickinger D., Pollard C., Wasow T. 1985. Structure-sharing in lexical representation. *Annual Meeting of the ACL Proceedings of the 23rd conference on Association for Computational Linguistics*: 262 - 267**.** Chicago, Illinois

Flickinger, D. 1987. *Lexical Rules in the Hierarchical Lexicon.* Ph.D. thesis, Stanford University

Fontenelle T., Adiaens G., De Braekeleer G. 1994. "The Lexical Unit in the Metal MT System", *Machine Translation*, Kluwer Academic Publishers. Vol. 9: 1-19

Friedman J. 1969. A computer system for transformational grammar *Communications of the ACM*. Vol. 12, Issue 6 : 341-348. ACM Press. New York, NY, USA.

Garside R., Leech G., Sampson G. 1987. *The Computational Analiysis of English*. Longman.

Gazdar G., Klein E., Pullum G. & Sag I. 1985. *Generalized Phrase Structure Grammar*. Cambridge, Massachussetts: Harvard University Press.

Gellerstam M. 1995. Brolexikon. *I Nordiske studier i leksikografi III*: 159-165. Reykjavík.

Goenaga P. 1980. *Gramatika bideetan*. Erein

Gojenola K. 2000 *Euskararen sintaxi konputazionalerantz. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan*. Doktore-tesia. UPV-EHUko Informatika Fakultatea.

González J., Ortiz D., Tomás J., Casacuberta F. 2004. A comparison of different statistical machine translation approaches for Spanish-to-Basque translation. III Jornadas en Tecnología del Habla. Valencia

Green G. 1989. *Pragmatics and Natural Language Understanding.* Hillsdale, N.J.: Lawrence Erlbaum.

Grishman R., Sterling J.: Acquisition Of Selectional Patterns. COLING 1992: 658-664

Gross M. (1975). *Méthodes en syntaxe. Régime des constructions completives*. Paris: Hermann.

Grover et al. 1993., 58

Grover, C., J. Carroll & Briscoe E. 1993. *The Alvey Natural Language Tools Grammar* (4th release), Technical Report No. 284. University of Cambridge.

Haan, P. 1984. Problem-oriented tagging of English corpus data. In Aarts, J. and Meijs, W. (arg.) *Corpus Linguistics*. Amsterdam: Rodopi.

Halliday M. 1991. Corpus studies and probabilistic grammar. In Aijmer and Altenberg 1991, pp 30-43.

Halliday M. & Hasan R. 1976. *Cohesion in English*. Londres: Longman.

Heid U. 1994. *On Ways Words Work Together -- Topics in Lexical Combinatorics* in Willy Martin et al., editor, Proceedings of the VIth Euralex International Congress 226--257. Amsterdam eingeladener Hauptvortrag.

Heid U. 1991. *Towards reusable lexical resources for natural language processing. Some proposals for linguistic knowledge representation* in Eleventh International Conference `Expert Systems and their Applications, Avignon, France. `Specialized Conference: Natural Language Processing and its Applications' EC2: 89 -- 101 Nanterre.

Hinkelman E.A. & Allen J.F. 1989. Two constraints on speech act ambiguity. *In Proceedings of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

Hirst G. 1987. *Semantic interpretation and the resolution of ambiguity.* Cambridge: Cambridge University Press, 1987.

Hopcroft J. and Ullman J. Introduction to Automata Theory, Languages and Computatuion. Addison-Wesley. 1979.

Hualde J.I. & Ortiz de Urbina J. (arg.) 2003. *A Grammar of Basque*. Mouton de Gruyter.

Hutchins W.J, Somers H. 1992. *An Introduction to Machine Translation*. Academic Press.

Ide N., Le Maître J., Véronis J. 1993. *Outline of a Model for Lexical Databases*. Information Processing and Management, vol. 29, no. 2, pp. 159-186.

Ide N. & Véronis J. 1993. *Extracting knowledge bases from machine-readable dictionaries: Have we wasted our time?* KB&KS WORKSHOP, Tokyo.

Ide N., Véronis J. 1998. Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, *Computational Linguistics*, vol. 24 n. 1, 1-40, 1998.

Ingria R. 1986. Lexical Information for Parsing Systems: Points of Convergence and Divergence, *Workshop "Automating the Lexicon"* (Grosseto).

Jensen K., Heidorn G., Richardson S. 1993. *Natural Language Processing: the PLNLP Approach*. Kluwer Academic Publishers, Boston.

Johansson S. & Stenström A-B. (arg.) 1991. *English Computer Corpora: Selected Papers and Research Guide*, Berlin: Mouton de Gruyter.

Johnatan S. 1988. *Machine Translation Systems*. Cambridge University Press.

Jurafsky D. Martin J. H. 2000. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.

Karlsson F. 1990. Constraint Grammar as a Framework for Parsing Running Text. In *Procs. CoLing'90*. In *Procs. 14th International Conference on Computational Linguistics*, ICCL, 1990.

Karlsson F., Voutilainen A., Heikkilä J. & Anttila A. (arg.). 1994. *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Karlsson F., Voutilainen A., Heikkilä J., Anttila A. (arg.) 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Berlin: Mouton de Gruyter.

Karp D., Schabes Y., Zaidel M., Egedi D. 1992. A freely available wide coverage morphological analyzer for English. Proceedings of COLING´92. Nantes. 950-954.

Karttunen L. and Beesley K.R. Two-Level Rule Compiler. Xerox ISTL-NLTT-1992-2. 1992.

Karttunen L. Constructing Lexical Transducers, Proc. of COLING´94, 406-411. 1994.

Karttunen L. Finite-State Lexicon Compiler. Xerox ISTL-NLTT-1993-04-02. 1993.

Karttunen L., Chanod J-P., Grefenstette G., Schiller A. 1997. *Regular Expressions For Language Engineering*. Natural Language Engineering.

Kasher A. (arg.) 1998. *Pragmatics: Critical Concepts.* Londres: Routledge.

Kelly E. & Stone P. 1975. Computer Recognition of English Word Senses. North-Holland, Amsterdam.

Kiraz G. 2000. Finite-State Morphology: Theory, Appications and Recent Developments. Tutorial in ANLP/NAACL'2000. Seattle.

Korta K. 2001. Begiratu zabala gaur egungo Pragmatikari. *Gogoa I-2,* 195-224.

Korta K. 1996. *Elkarrizketaren eredu baterantz: asmoa, ekintza, komunikazioa / Hacia un modelo del diálogo: intención, acción, comunicación*. (Elebitan: euskara / gaztelania.) Donostia: ILCLI eta UPV-EHU Argitarapen Zerbitzua.

Koskenniemi K. 1983. *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki. Publications nº 11.

Koskenniemi K. 1985. Compilation of Automata from Morphological Two-level Rules. University of Helsinki, Publication nº 15.

Koskenniemi K., Tapanainen P., Voutilainen A. 1992. *Compiling and using finite-state syntactic rules.* COLING'92, Nantes.

Kukich K. 1992. Techniques for automatically correcting word in text. ACM Computing Surveys, vol.24, No. 4, 377-439.

Kytö M., Rissanen M., Wright S. (arg.) 1994. *Corpora across the Centuries*, Amsterdam, Rodopi.

Laka I. 1998. *A Brief Grammar of Euskara, the Basque Language*. HTMLko dokumentua. Euskararako Errektoreordetza, Euskal Herriko Unibertsitatea. http://www.ehu.es/grammar/index.htm

Lawler J., Aristar H. 1998. *Using Computers in Linguistics. A practical Guide*. Routledge.

Leech G. 1993. Corpus annotation schemes. *Literary and Linguistic Computing* 8(4): 275-81.

Leech G. & Fallon R. 1992. Computer corpora - what do they tell us about culture?. *ICAME Journal* 16: 29-50.

Lenders W. 1990. Semantische Relationen in Wörterbuch-Einträgen. Eine Computeranalyse des DUDEN-Universalwvrterbuchs. In: *Burkhard Schaeder/Burghard Rieger (Hg.): Lexikon und Lexikographie*. Hildesheim. 92-105

Levin B. 1993. *English verb classes and alternations.* The University of Chicago Press.

Lewis D, & Sparck Jones K. 1996. Natural Language Processing for Information Retrieval. Communications of ACM Vol.39, Num.1.

Litman D.,-Hirchberg J. 1990. Disambiguating cue phrases in text and speech *Proceedings of the 13th conference on Computational linguistics*. Volume 2.

Marcu D. 2000. The Theory and Practice of Discourse Parsing and Summarization. MIT Press.

Marcus M., Santorini B. 1991. Building very large natural language corpora: the Penn *Treebank*. CIS report, University of Pennsylvania.

Miller G. 1990. Five papers on WordNet. Special Issue of International Journal of Lexicography 3 (4).

Mindt D. 1991 "Syntactic evidence for semantic distinctions in English", in Aijmer and Altenburg 1991: 182-96.

Montague, R. 1970. Pragmatics and intentional logic. *Synthese 22*: 68-94.

Montague R. 1973. The proper treatment of quantification in ordinary English. In Jaakko Hintikka ed. *Approaches to Natural Language: Proceedings of the 1970 Stanford Workshop on Grammar and Semantics*, pages 221-242. D. Reidel Publishing Co., Dordrecht, Holland. Reprinted in *Formal Philosophy*, by Richard Montague, Yale University Press, New Haven, CT, 1974, pp. 247-270.

Montague R. 1974. The Proper Treatment of Quantification in Ordinary English, in Formal Philosophy: Selected Papers of Richard Montague, ed. by Richmond Thomason, Yale University Press.

Myers G. 1991. "Pragmatics and corpora", talk given at *Corpus Linguistics Research Group*, Lancaster Univeristy.

Neff M. & Boguraev B. 1989. Dictionaries, dictionary grammars and dictionary entry parsing. *Proc. of 27th ACL.,* Vancouver, pp. 91-101.

O'Connor J. & Arnold G. 1961. *Intonation of Colloquial English*. Londres: Longman.

Pustejovsky J. 1991. The Generative Lexicon: A Theory of Computational Le~:ical Semantics. MIT Press, Cambridge.

Pustejovsky J. 1995. *The Generative Lexicon*. Cambridge, Londres: MIT Press.

Ravin Y. 1990. Disambiguating and interpreting verb definitions. *Proceedings of the 28th conference on Association for Computational Linguistics*. Pittsburgh, Pennsylvania. Pages: 260 - 26

Ravin Y. & Leacock C. 2000. *Polysemy. Theoretical and Computational Approaches*. Oxford University Press.

Daudé J., Padró L. & Rigau G. 1999. Mapping Multilingual Hierarchies using Relaxation Labelling, Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99). Maryland, United States.

Atserias J., Castellón I., Civit M. & Rigau G. 1999. Using a Diathesis Model for Semantic Parsing, 1th Venezia per il Trattamento Automatico delle Lingue (VExTAL'99). Venice, Italy.

Roche R., Schabes Y. 1997. *Finite-State Language Processing*. MIT Press.

Sampson G. 1987. *Evidence against the 'Grammatical/Ungrammatical' Distinction.* Corpus Linguistics and Beyond, W. Meijs editorea, Rodopi, Amsterdam.

Schmidt K. M. .1993. *Begriffsglossar und Index zu Ulrichs von Zatzikhoven Lanzelet*. Tübingen: Niemeyer.

Sedelow S. & Sedelow W. 1969. *Categories and procedures for content analysis in the humanities*. In Gerbner G., Holsti O. R., Krippendorff K., Paisley W. J. & Stone P. J. (eds) *The Analysis of Communication Content*. New York. John Wiley.

Shieber S.M. 1986. *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes 4, Stannford.

Shieber 1987, 31

Shieber M. 1987. An Introduction to Unification-Based Approaches to Grammar. *CSLI Lecture Notes*, no. 4.

Sparck Jones K., Galliers J.R. 1996. *Evaluating Natural Language Processing Systems*. Lecture Notes in Artificial Intelligence, Springer.

Sperberg-McQueen C. M. & Burnard L. 1994. *Guidelines for Electronic Text Encoding and Interchange (P3)*. Chicago and Oxford: Text Encoding Initiative.

Stenstöm, A-B. 1984a. "Discourse items and pauses", Paper presented at *Fifth ICAME Conferenc*e, Windermere. Abstract in *ICAME News* 9 (1985): 11.

Stenstöm, A-B. 1984b. *Discourse tags*. In Aarts J. & Meijs W. (eds) *Corpus Linguistics*. Amsterdam: Rodopi.

Tzoukermann E. & Liberman M. 1990.A finite-state morphological processor for Spanish. Proc. of COLING-90, Helsinki, vol.3, 277-281.

Urkia, M. 1997. *Euskal morfologiaren tratamendu informatikorantz*. Doktore-tesia, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Voutilainen, A., Tapanainen P. 1993. *Ambiguity resolution in a reductionistic parser.* EACL'93, Utrecht.

Voutilainen, A. 1994a. *Three studies of grammar-based surface parsing of unrestricted English text.* Ph.D. thesis. University of Helsinki. Publications nº 24.

Voutilainen, A. 1994b. *Designing a parsing grammar.* University of Helsinki. Publications nº 22.

(Walker eta Amsler, 1986) Walker D. & Amsler R. 1986. The use of machine-readable dictionaries in sublanguage analysis, *Analysing language in restricted domains*, Lawrence Erlbaum, Hillsdale, NJ. pages: 69-84.

Woods W. A., Kaplan R. A. & Nash-Webber B. 1972. The Lunar Sciences Natural Language Information System: Final Report: BBN Report # 2378. Bolt Beranek and Newman Inc., Cambridge, MA.

11 Glosategia

algoritmoa
Problema bat urrats kopuru finituan ebazteko ekintza multzo esplizitua.

alomorfoa
Bi morfema alomorfo direla esaten da informazio morfologiko bera eta forma kanoniko bera dutenean. Horrela, gaztelaniaz "jug" eta "jueg" izan daitezke sistema baten sarrera alomorfoak, biek adierazten baitute "jugar" aditzaren erroa. Ez da gomendagarria asko erabiltzea, mantentze-lana zailtzen dutelako, baina batzuetan haien erabilerak sistemaren garapena errazten du.

analisi semantikoa
Analisi semantikoaren helburua esaldiaren esanahia lortzea da, hau da, bere edukiaren errepresentazio kontzeptuala sortzea. Horretan, esaldiaren esanahia egitura formal baten bidez adierazi beharko da, eta horrelako adierazpideei *esanahi-adierazpide* deituko diegu.

analizatzaile morfologikoa
Hitzak zati morfologikotan banatu eta morfemen arteko loturak bideratzen dituen tresna.

anbiguotasuna
Morfosintaxiaren mailan, forma batek analisi-lerro bat baino gehiago duenean gertatzen da anbiguotasuna. Hau da, hitza testuingururik gabe analizatzen denean, kategoria, azpikategoria, kasua... bat baino gehiago dituenean. Hizkuntzalaritzan *sinkretismo* ere deitua izan da eta kasu tipikoa absolutibo plurala eta ergatibo singularraren artean gertatzen dena da (adibidez, *emakumeak* hitzean gertatzen dena). Sintaxiaren mailan, analisi-lerro bakoitzak funtzio sintaktiko bat baino gehiago duenean gertatzen da anbiguotasuna (adibidez, absolutibo singularra subjektu, objektu eta predikatibo izan daiteke). Semantika mailan ere, esate baterako, hitz batek adiera bat baino gehiago dituenean.

automata
Eredu matematikoa, egoera multzo bat eta egoera-trantsizioen baldintzak definitzen dituen arau multzo batek osatua.

baterakuntza
Ezaugarri-egituren artean definitzen den eragiketa. D' eta D" egituren baterakuntzaren emaitza (baldin badago) beste egitura bat da: D' eta D" egitura bien hedapena den egitura orokorrena. Hau da, D' egituran dauden ezaugarriak emaitzan egongo dira, D" egiturarenak ere bai, eta emaitzan ezaugarri horiek dituzten balioak bateragarriak dira D' eta D" egituretan zeuzkatenekin. Adibidez:

$$D' = \left[\begin{array}{lll} \text{komunztadura:} & \text{numeroa:} & \text{singularra} \\ \text{subjektua:} & \text{numeroa:} & \text{singularra} \end{array}\right]$$

$$D'' = \left[\begin{array}{lll} \text{subjektua:} & \text{pertsona:} & 3 \end{array}\right]$$

$$D' \cup D'' = \left[\begin{array}{lll} \text{komunztadura:} & \text{numeroa:} & \text{singularra} \\ \text{subjektua:} & \text{numeroa:} & \text{singularra} \\ & \text{pertsona:} & 3 \end{array}\right]$$

baterakuntza-formalismoak
Hitzek osatzen dituzten unitate handiagoak, ezaugarri-egituretan oinarritutako herentzia-mekanismoez lortzen dituzte.

berrerabilgarritasuna
Bi eratara uler daiteke: batetik, hasiera batean giza erabilerarako pentsatuta zeuden baliabide lexikalak LNPrako *berrerabili* ahal dira; bestetik, LNPrako tresnak eta hizkuntza-baliabideak esaterako, ondo

diseinatuz gero, hainbat aplikaziotan *berrerabil* daitezke, behin eta berriz informazio bera antolatzen ibili gabe.

bitestua (corpus paraleloa)
Testu bera bi hizkuntzatan duten corpusak dira.

***botton-up* analisi sintaktikoa**
Testuingururik gabeko gramatika batek sortutako unitateak (perpausak, etab.) analizatzeko estrategia, zuhaitz sintaktikoa eraikitzen duena hostoetako adabegietatik hasi eta errorantz jarraituz.

chunk (ikus **zati)**

corpus
Corpusak, testu-biltegiak dira, eta metodo enpirikoetarako aukera ematen dute. Linguistika enpiriko eta deskriptiboaren oinarriak dira.

corpus eleaniztunak
Zenbait parametroren arabera ezaugarri komunak agertzen dituzten bi hizkuntzatako edo gehiagotako testu-bildumak. Bi mota nagusi bereizten dira: corpus paraleloak eta corpus konparagarriak.

corpus gordina
Inongo markaketarik gabeko corpusak dira.

corpus konparagarriak
Bi hizkuntzatako hainbat testu biltzen dituzten corpusak dira. Testu horiek ezaugarri (gaia, urtea, mota...) berekoak dira, baina ez dira bata bestearen itzulpena.

corpus lerrokatuak
ik. corpus parekatuak

corpus markatua (corpus etiketatua)
Informazio linguistikoarekin aberastutako corpusak. Hainbat marka dituzten corpusen erabilera markarik gabekoena baino askozaz handiagoa da. Testu markatuetan, testu batek ezkutuan duen informazioa agerian uzten da, hainbat markaketa-prozesuren bitartez kodetuz, eta gero horiek baliatzen dira hainbat zereginetarako.

corpus paraleloak (bitestua)
Elkarren itzulpen diren testuez osatuak. Batzuetan, gerta daiteke itzulpena erabatekoa ez izatea, eta testu batetik bestera pasarteak falta izatea, edo informazioa erabat berdina ez izatea; horrelakoei corpus paralelo ‘zaratatsuak’ esaten zaie (‘noisy parallel texts’).

corpus parekatuak (corpus lerrokatuak)
Corpus paraleloak (edo bitestuak) dira, baina hizkuntzen arteko unitate linguistikoak elkarri lotuta daude. Esaldiz esaldi parekatutako corpusetan bata bestearen itzulpena diren esaldiak markatzen dira. Perpausaz gain badira beste parekatze maila posibleak: hitzak, terminoak, lokuzioak, kolokazioak, entitateak...

chart
Analisi sintaktikoa eginez ezagutu diren osagai sintaktiko posible guztiak, sinple eta konposatuak, biltzeko erabiltzen den taula.

datu-base lexikal (DBL)
Lexikoaren gainean biltzen den ezagutza mota gehienbat gramatikala denean (kategoria, azpikategoria, morfotaktika...), *datu-base lexikal* (DBL) terminoa erabiltzen da.

desanbiguatu/desanbiguazioa
Anbiguotasuna gertatzen denean (ik. lehenengo glosa), testuinguruari begiratzen zaio hitz batek aukeran dituen analisi-lerroen artean egokiena zein den jakiteko. Testuinguru jakin horri ez dagokion analisi-lerroa kentzea ala dagokiona besterik ez uztea da desanbiguatzea. Modu berean, lerro horri ez dagokion funtzio sintaktikoa kentzea litzateke ala dagokiona besterik ez uztea. Murriztapen Gramatika erabiltzen bada desanbiguatzeko, erregelen bitartez egiten da.

desanbiguazio morfosintaktikoa
Hitz-forma baten osaera morfologiko posibleetatik, testu zatian dagokion interpretazioa esleitzeko beharrezkoa den prozedura.

diakritikoak
Informazioa morfofonologikoa modu kriptikoan eman eta forma kanonikoa desitxuratzen duten sinboloak dira. Salbuespenen detekzioa eta erregelen aplikazioa kontrolatzea da haien helburua, baina ahal denean, lexiko-itzultzaileetan adibidez, saihestu behar dira bestelako informazio morfologikoa erabiliz.

dokumentuen berreskurapena (IR, *Information Retrieval*)
Biltegiratutako dokumentu-biltegi batetik dokumentu espezifikoak eskuratzeko metodo eta teknikak.

dokumentuen bideratzea (*Routing*)
Aplikazio honetan dokumentua sailkatzeaz gain, dokumentuaren kategoriari dagokion helbidera edo sailera bidaltzen du dokumentua aplikazioak, dokumentuari postratamendu espezifiko bat emanez.

dokumentuen iragaztea (*Filtering*)
Dokumentuen ezaugarri batzuk detektatu eta horren arabera dokumentu bera baztertu ala onartu egiten dute horrelako sistemek. Aplikazio honen adibide tipikoa posta elektronikoko *spam*-mezu guztiak detektatzea eta automatikoki alde batera uztea da.

dokumentuen laburpena edo laburpen automatikoa (*Summarization*)
Aplikazio honen zeregina dokumentuen laburpena automatikoki egitea da. Bi eratara bidera daiteke. Modu errazena da testu zati edo esaldi esanguratsuenak hautatzea. Modu zaila erabiltzen denean, aldiz, ideia nagusiak detektatu, integratu eta testu berri bat sortzen da.

dokumentuen multzokatzea (*Clustering*)
Clustering egiten denean, aldez aurretik ez daude definituta sailkapenerako kategoria posibleak. Abiapuntuan, hainbat dokumentu dauzkagu, eta bukaeran dokumentu horiek guztiak sailkatuta, haien arteko antzekotasunen arabera. Jakin beharko da geroago interpretatzen zergatik proposatu diren multzo horiek, zer adierazten duten azpimultzo horiek.

dokumentu-salkatzaileak
Sailkatze-sistemak oso baliagarriak dira dokumentu ugari kategoria multzo txiki baten arabera sailkatu behar izanez gero. Adibidez, hainbat albiste banatzea *kirola, nazioartekoa, kultura* eta *herrikoa* kategorien artean. Edota, Yahoo bezalako bilatzaile baten kasuan, web orri berri bat detektatzen duenean, zehaztea zein gaitan kokatu beharko den.

DTD (*Document Type Definition*)
Dokumentu motaren definizioa. DTDn dokumentuaren mota formalki definitzen da, hots, bere osagaiak eta egitura esplizituki adierazten dira XML, SGML markatze-lengoaietan.

egoera finituak
Grafoetan oinarritutako konputazio-eredu sinplea eta, ondorioz, azkarra programaren abiaduraren aldetik. Hizkuntza-teknologietan egoera finituetako eredua kontrajartzen zaio *baterakuntza-mekanismoa*ri, azken hori konplexuagoa eta, ondorioz, motelagoa izanik. Sintaxiaren eredu klasikoak *baterakuntza*n oinarritzen dira, egoera finituekin ezin baitira fenomeno sintaktiko guztiak adierazi. Dena den, egoera finituetan oinarritutako tresnak erabiltzen dira gaur egun tratamendu sintaktiko partziala burutzeko.

elkarketa
Hitza lema batek baino gehiagok osatzen dutenean.

elkarrizketa-sistemak
Elkarrizketa-sistemek sistema informatiko batekin harremanetan jartzea ahalbidetzen dute, hartara, informazioa eskuratu edota trukeak egin ahal izateko. Horretarako, analisia eta sorkuntza dute integraturik, eta, horrez gain, elkarrizketa kudeatzeko modulu bat.

eratorpena
Hitza lema batek eta eratorpen-atzizki batek (edo gehiagok) osatzen dutenean.

espektrograma
Soinuen ezaugarrien adierazpen grafikoa. Ardatz horizontalak denbora adierazten du, eta ardatz bertikalak maiztasuna (Hz). Gris mailak (batzuetan koloreak erabiltzen dira) maiztasun baterako soinuak duen energia erakusten du.

etiketatzailea
Hitzari testu zatian dagokion interpretazioa esleitzen dion tresna, bere baitan prozesu desanbiguatzaileak baliatzen dituena. Horretaz gain, funtzio sintaktikoak esleitzeko ahalmena du, eta testua sintaktikoki etiketatzea ahalbidetzen du. Hala, etiketatzaileak analizatzaile sintaktiko partzialak dira neurri handi batean.

etiketatze (*tagging*)
Zenbait markaketa linguistiko, hala nola hitzei kode bereziak atxikitzea haien zenbait ezaugarri adierazteko, *markaketa* gisa baino gehiago *etiketatze* (*tagging*) gisa ezagutzen dira; eta ezaugarriei egokitzen zaizkien kodeei *etiketa* (*tag*) esaten zaie. Etiketatzea zenbait kontu markatzeko erabiltzen da. Eta horregatik maila desberdinetako etiketatzeak daude.

ezagutza-base lexikal (EBL)
Hitz eta adierei buruzko informazioa duten lexikoiak dira. EBLen ezaugarri garrantzitsuena herentzia izaten da, adierak klase/azpiklase hierarkien inguruan antolatzen dira eta.

ezaugarri-egitura (*feature structure*)
Hizkuntz ezagutza adierazteko modu bat. Hainbat ezaugarri biltzen da egitura batean, ezaugarri bakoitza bere balioarekin. Balioak ere ezaugarri-egiturak izan daitezke. Adibidez, ondoko egitura honetan bi ezaugarri definitu dira: *cat* eta *head*. Azken ezaugarri horren balioa, era berean, beste ezaugarri-egitura da, *form* eta *subject* ezaugarriak definituta dituena, hain zuzen ere.

cat: S				
head:	form: finite			
	Subject:	agreement:	number: sing person: 3	

flexioa
Lemari informazio morfosintaktikoa ematen dion atzizkia.

formalismo
Ezagutza linguistikoaren ereduen oinarrian dauden erregelak irudikatzeko baliabidea.

gainsorkuntza
Deskribapen morfologiko batean hizkuntzaren forma okerrak analizatzen edo sortzen direnean, deskribapena gainsortzailea dela esaten da. Oso zaila da estaldura osoa lortzea eta gainsorkuntza erabat baztertzea, batez ere eratorpenean eta elkarketan, estaldura osatzeko generalizazioak egin behar direlako, gainsorkuntza eraginez.

galdera-erantzuneko sistemak (*Question Answering*)
Interfaze berezi hauen helburua da erabiltzailearen galderentzako erantzunak aurkitu eta itzultzea. Galderak lengoaia naturalez egiten dira eta erantzunak ere lengoaia naturalez eman behar dira.

gizakiak lagundutako ordenagailu-itzulpena
Gizakiak lagundutako ordenagailu-itzulpenean, ordenagailuak egiten du itzulpena eta gizakiak orrazten du ondoren.

GPSG (*Generalized Phrase Structure Grammar*)
Baterakuntzan oinarritutako gramatika formalismoa.

hizketa-ezagutza
Ahots-seinalea duen mezua interpretatzea.

hizketa jarraitua
Hitzen arteko isilunerik gabeko hizketa.

hizketaren analisia (*Speech recognition*)
Sintesiaren kontrako teknika dugu eta honetan datza: seinale akustiko bat sistema informatiko batek interpreta dezakeen errepresentazio sinboliko bihurtzean.

hizketaren sintesia (TTS, *Text to Speech*)
Ahozko mezuak testu batetik automatikoki sortzean datza hizketaren sintesia. Ahotsa sistema informatikoak berak sortuko du aurretik zehaztutako datu multzoa edota erregelak baliatuz. Sintesirako hainbat estrategia daude.

hizketaren tratamendua
Hizkuntza mintzatuaren azterketa konputazionalaz arduratzen den arloa. Hizketaren tratamenduaz aritzerakoan, bi sistema nagusi garatzen dira: hizketaren ezagumendua edo analisia (*Speech Recognition,* SR), eta sintesia edo sorkuntza (*Text to Speech,* TTS).

hizkuntza eranskaria
Informazio morfosintaktikoko hainbat morfema independente kateatzen dituen hizkuntza.

hizkuntza flexiboa
Informazio morfosintaktikoko morfemak ezin bereiz daitekeen morfema bakarrean kateatzen dituen hizkuntza

hizkuntzalaritza informatikoa
Hizkuntzaren azterketarako lagungarriak diren programak lantzeaz arduratzen den arloa.

hizkuntzalaritza konputazionala (HK)/(*Computational Linguistics,* CL)
Ikuspegi abstraktuago batetik ekiten dio hizkuntzaren modelizazioari ordenagailuek hizkuntza uler dezaten.

hizkuntza-teknologia
Hizkuntzalaritza teorikoa eta praktikoa, informatika eta adimen artifiziala biltzen dituen diziplina, edo, bestela esanda, alde teorikoa eta ingeniaritzari dagokiona ere bai.

hiztegi datu-baseak (HDB)
Hiztegia euskarri elektronikoan badago, hiztegi datu-basetzat har daiteke.

hiztegi ezagutza-baseak (HEB)
HEBek hiztegietatik erauzitako informazioa jasotzen dute. Erauzitako informazioen artean, EBLetan bezala, hemen ere, adieren hierarkiak dira aipagarriak.

hiztun-ezagutza (*Speaker recognition*)
Hiztuna nor den bereizteko metodoa.

HPSG (*Head Phrase Structure Grammar*)
Baterakuntzan oinarritutako gramatika formalismoa. Guneak zuzendutako egitura sintagmatikoen gramatika.

HTML
HyperText Markup Language.

informazio erauzketa (*Information Extraction,* IE)
Testuetatik edo hizketatik informazio adierazgarria automatikoki ateratzea.

ingeniaritza linguistikoa (IL)(*Language Engineering,* LE)
Hizkuntzari buruzko ezagutza batez ere aplikazioetara eta produktu komertzialetara zuzenduta dago. Hizkuntza ezagutzeko, ulertzeko, interpretatzeko eta sortzeko gai diren sistema informatikoak garatzea du jomuga.

interfazeak
Gizakiaren eta makinaren arteko elkarrekintzan laguntzeko sistemak. Bi sistema mota dira: komunikazioa testu idatziaren bidez bideratzen dutenak eta hizketaren bidez egiten dutenak.

interlingua
Hizkuntzatik hizkuntzara itzultzeko bitarteko lengoaia neutral bat erabiltzen duen estrategia. Honek inolako transferentziaren beharrik ez izatea ekartzen du; aldiz, analisia eta generazioa dira oinarriak.

interpretazio semantikoa
Testuingurua kontuan hartu gabe, esaldiaren esanahi abstraktua lortzen duen analisi-fasea. Forma logiko baten bitartez adierazten da esaldiaren esanahia.

itzulpen zuzena
Hizkuntzatik hizkuntzara itzultzeko erabiltzen duen estrategia zuzena. Horretan ez da transferentziazko modulurik ez eta hizkuntzak errepresentatzeko lengoaia neutral komunik behar. *String*en ordainak baino ez dira ematen. Baliagarria da jatorrizko hizkuntza eta xede-hizkuntza oso antzekoak direnean.

jarraitze-klase
ik. morfotaktika

klitikoa
ik. morfema

konposagarritasuna
Interpretazio semantikoa prozesu konposizionala da. Hau da, elementu baten esanahia bere osagaien esanahiak konbinatuz lortzen da.

lema
ik. morfema

lematizatzailea
Testu bateko hitz-forma bakoitzeko lema zein den definitzen duen tresna konputazionala.

Lengoaia Naturalaren Prozesamendua (LNP)/(*Natural Language Processing,* NLP)
Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloari Lengoaia Naturalaren Prozesamendua (LNP) esaten zaio, eta, batez ere, erabiliko diren teknika informatikoei erreparatzen dio: algoritmoak, konpilatzaileak, estrategiak, etab.

lexikoia
LNPren arloan informazio lexikalaren biltegiei edota hiztegiei erreferentzia egiteko erabiltzen den terminoa.

LFG (*Lexical Functional Grammar*)
Baterakuntzan oinarritutako gramatika-formalismoa. Gramatika lexiko funtzionala.

Linguistika Konputazionala (LK)/ (*Computational Linguistics,* CL)
Ikuspegi abstraktuago batetik ekiten dio hizkuntzaren modelizazioari ordenagailuek hizkuntza uler dezaten. Hau da, hizkuntza formalizatzen du ordenagailuek ulertu ahal izateko moduan.

***mappings*/mapaketa-erregelak/ islapen-erregelak**
Murriztapen Gramatikan erabiltzen diren erregelak dira eta gramatikaren atal bat osatzen dute. Hitzei beren etiketa morfologiko edo sintaktiko posibleak gehitzeko erabiltzen dira.

Markov Eredu Ezkutua (MEE)
Eredu estruktural estokastikoa. Egoera finituen makina probabilistikoa.

MEEren topologia
Ereduaren egoera kopurua eta egoeren arteko trantsizioak.

meta-datua
Datuari buruzko datua edo informazioa. Adibidez, “Mikel” datua da eta “Pertsonaren izena” meta-datua da, datuari buruzko informazioa delako, kasu honetan esanahia.

morfema
Morfologiari begirako unitate linguistikoa. Bi mota bereizten dira: *lema* eta *erroak*, hitza osatzeko ezinbestekoak direnak eta lexikoetan eta hiztegietan agertzen direnak, eta *hizkiak*, aurrekoen modifikatzaileak direnak. Azken horiek 3 motatakoak izan daitezke: hitzaren hasieran agertzen diren *aurrizkiak*, hitzaren bukaeran kokatu ohi diren *atzikiak* eta hitz barruan txertatzen diren *artizkiak*.

morfofonologia
Morfemak biltzean gertatzen diren aldaketen definizioa. Adibidez, euskaraz askotan irakurtzen den erregela morfofonologiko bat hau da: *a*+*a* = *a*. Horrek esan nahi du: *a*-z bukatutako morfema bati *a*-z hasitako bat lotzen zaionean *a* bakarra geratzen da. Aldaketa hauen iturria fonologikoa da batzuetan baina beste batzuetan ortografikoa ere izan daiteke.

morfotaktika
Morfemen arteko konbinazio posibleen definizioa. Behin morfemak definituta daudela zehaztu egin behar dira haien arteko konbinazio posibleak. Sistema batzuetan morfotaktikaren definizioa *paradigma* bitartez egiten da. Bi mailako morfologian morfotaktika definitzeko jarraitze-klase bat esleitzen zaio morfema bakoitzari, jarraitze-klasea ondoren joan daiteke morfema multzoaren identifikazioa delarik. Adibidez, euskaraz adjektiboen jarraitze-klasean graduatzaileak egongo dira (*-ago*, *-egi*, *-en*) baina izenenean ez.

MRD (Machine Readable Dictionary)
Euskarri magnetikoan gordetzen den hiztegia. Hiztegi elektronikoa.

Murriztapen Gramatika
Hitz bakoitza bera bakarrik agertzen denean analizatzeko dituen aukera morfologiko eta sintaktiko guztietatik abiatuz, erregelen bidez eta testuinguruko beste hitzak kontuan hartuz aukera horiek murrizten dituen formalismoa. Aldez aurretik etiketa morfologiko eta sintaktikoak gehitzen zaizkie hitzei, bere interpretazio posible guztiak kontuan hartuz.

OCR (*optical character recognition)*
Makina batek paper batean inprimatuta dauden karakteretan testua ezagutu, eta kodetu egiten duen prozesua, geroago testu hori konputazionalki tratatu ahal izateko.

ontologiak
Mundu errealaren kontzeptualizazioak dira, hitzekin izendatzen ditugun kontzeptuak modu hierarkikoan antolatuta, mundu errealari buruzko inferentziak egiteko gaitasuna dutenak.

parametrizazioa
Hizketaren seinaletik ezaugarri garrantzitsuenak atera.

PATR
Baterakuntzan oinarritutako gramatika formalismo sinple eta oinarizkoa.

pitch-aldaera
Esaldi osoaren doinua, galdera, baiezpen edo beste ideiak adierazteko.

pragmatika
Pragmatika testuinguruari dagokion informazioaz arduratzen da; hau da, berez linguistikoa ez den, eta igorpen linguistikoen prozesamenduan eta interpretazioan eragina duten informazioez arduratzen da. Bi atal bereiz daitezke hor: diskurtsoaren ezagutza eta munduaren ezagutza.

Prolog
Lehen mailako predikatu-logikan oinarritzen den programazio-lengoaia. Hainbat erabilera izan ditu hizkuntza-aplikazioetan. Euskararen PATR analizatzaile sintaktikoa SICStus Prolog lengoaia erabiliz inplementatu da.

prosodia
Ahotsaren pitch, bolumena, iraupena eta erritmoaren aldaketak.

SGML (*Standard Generalized Markup Language*)
Testuak markatzeko lengoaia estandar eta orokorra. Gaur egunean XML erabiltzen da, SGMLren ezaugarriak eta berri batzuk batzen dituena.

sintaxi partziala
Sintaxi partzialak analisi tradizionalaren informazioaren zati bat, ez guztia, aztertzen du. Helburu ez da perpaus osoa analizatzea zuhaitz bakarra lortuz, sintaxi partzialean perpausaren osagaiak identifikatzea aski da (izen-sintagmak, preposizio-sintagmak eta aditz-kateak batez ere). Sintaxi partzialean erabiltzen diren teknikek fidagarritasuna eta sendotasuna dute helburu, sakontasuna eta osotasuna neurri batean galduz.

sistema eraikitzaileak
Hitzetatik abiatuta, beraien konbinazioaz bakarrik esaldiaren azken egitura eraikitzen saiatzen dira, era determinista alferrikako aukerak sortu gabe.

sistema murriztaileak
Lexikoiko informazioa erabiliz, esaldia analizatzeko aukera posible guztiak sortzen dituzte, eta gero, gramatikariaren lana, perpauseko testuingurua aintzat hartuz aukera onartezinak baztertzea da.

soinua
Airearen aldaera-seriea da.

sorkuntza
Sorkuntza analisiaren kontrako prozesua dugu. Analisiaren eginkizun nagusia testuaren formatik testu horren errepresentazio abstraktura iristea da; sorkuntzan, aldiz, errepresentazio abstraktu batetik testura iristen gara.

TEI (*Text Encoding Initiative*)
Testuak kodetzeko ekimena. Hainbat testu mota (prosa, poesia, hiztegiak…) kodetzeko estandarrak proposatzen dituena.

teknika probabilistikoak
Ezagutza linguistiko minimoarekin (corpus etiketatuak batez ere) datu probabilistikoak lortzen dituzte hizkuntza-teknologian erabil daitezkeenak.

testu etiketatuak
Testu-fitxategiak dira, baina testuaren egitura islatzeko markez hornituta daude.

testuingururik gabeko gramatikak (*Context Free Grammar*)
Hizkuntzak konputazionalki aztertzeko gaitasuna duen gramatika-formalismoa. Erregeletako ezker aldean bukaerakoa ez den sinbolo bat azaldu behar da, eta eskuin aldean bukaerakoak diren edo ez diren hainbat sinbolo.

testu-mehagintza (*text mining*)
Datu multzo baten baitan egiten den azterketa, testutik korrelazioak eta informazioak atzemateko.

testu-sorkuntza automatikoa
Ordenagailu barruan dauden datu konplexuetatik abiatuz (inprimakiak, datu kodetuak edo zenbakizko formatuan dauden informazioak...), datu horien edukia azalduko zaio erabiltzaileari bere hizkuntzan.

***top-down* analisi sintaktikoa**
Testuingururik gabeko gramatika batek sortutako hitzak analizatzeko metodoa, zuhaitz sintaktikoa goitik behera eraikitzen duena errotik hasi eta adabegietarantz jarraituz.

transferentzia
Hizkuntzatik hizkuntzara itzultzeko transferentziazko moduluak erabiltzen dituen estrategia. Transferentzia maila desberdinetan egin daiteke. Normalean sintaktikoan edo semantikoan egin ohi da.

Transduktoreak
Bi sarrera hartzen dituzten automata bereziak. Egoera finituko tekniketan erabiltzen dira, batez ere morfologia lantzeko.

treebank
Sintaktikoki etiketatutako corpusa.

XFST
Osagai morfologiko eta sintaktikoak ezagutzeko Xerox-ek sortu duen tresna. Testuingururik gabeko gramatika baino sinpleagoak diren automatak erabiltzen dira hemen. Nonbait ahalmen apalagoa dute, baina oso azkar ibiltzen dira. Azaleko sintaxia nahikoa denean, aukera egokia izan daiteke automata hauek.

XML
eXtensible Markup Language

XSL
eXtensible Stylesheet Language

zarata
Hizketa-seinalearekin egoten diren gainerako soinuak.

Zati (*chunk*)
Testu bateko osagai sintaktiko sinpleak (izen-sintagmak, preposizio-sintagmak, aditz-lagunak, aditz-kateak...). Perpaus oso-osoen analisia lortzeko asmorik gabe, zatiak bereizteko tresnak ere sortu izan dira hizkuntza-teknologiako zenbait aplikazio garatzeko baliagarriak direlako.

12 Aurkibide alfabetikoa