LENGOAIA NATURALAREN ORDENADORE BIDEZKO APLIKAZIOAK

E.Agirre, I. Alegria, X. Arregi, X. Artola,
A. Díaz de Ilarraza, K. Sarasola eta J.R. Zubizarreta

Informatika Fakultatea
649 Postakutxa, 20080 Donostia

1. Sarrera

Inprentaren sorkuntzak hizkuntzaren tratamendua eta zabalkuntza irauli zuen moduan, mende honetan sortu den ordenadoreak horren pareko iraultza dakar. Etxeparek ospatu zuen euskarazko lehenengo liburuaren inprimatzea, baina oraindik ezin aurkituko dugu ordenadore bidez lehen testua aztertzea edo sortzea ospatuko duen Etxepare modernorik, nahiz eta oinarrizko zenbait lan euskararentzat egokitu diren.

Testu–prozesaketarako baliabide berria den ordenadoreak, gaur egun erraztasun handiak eskaintzen ditu testuak kopiatu, zabaldu eta zuzentzeko, eta baita mila formatu edo itxura desberdinetan aurkeztu ahal izateko ere. Baina testu–edizioko baliabide horiek baino askoz ere lan bitxiagoak aztertzen dituzte puntako ikertokitan, non helburu berriak hizkuntza ulertu eta sortzean kokatzen bait dira. Gizakien arteko komunikaziorako erabiltzen den lengoaia idatzi zein ahoskatzeko, tresna lagungarriak eraikitzen ari dira. Ordenadorearen betebeharra zuzentzeko behar den zehaztapena bigunago bihurtuko duten bitarteko linguistikoak, proiektu ez eta errealitate dira gaur egun. Bitarteko hauei esker, programak eraikitzeko programazio- –lengoaiak ezagutzea ez da derrigorrezkoa izango.

Dena dela, azken urteotako lorpen hauek ez dira guztiz orokorrak. Esaldi arrunt batzuk uler ditzakete honelako sistemek, baina beti asma liteke ulertezin suertatuko zaien esaldiren bat. Aplikazio zehatz eta mugatu bati buruzko esaldiak erabil ditzakete, baina mintzagai horretatik kanpo emaitza eskasak lortuko dira.

Bost urteko edozein umek hitz egiten eta ulertzen ondo dakienez, hizkuntza erabiltzea lan erraza dela pentsatzen dugu, baina hori ez da horrela. Lengoaia ulertzea oso prozesu konplexua da eta gaur eguneko ordenadoreak urrun ikusten ditu giza adimenaren ahalmen linguistiko eta orokorrak.

Telebistako kirol–komentarista baten azalpena, 093 telefonoan zer ordu den esatean duen leloa eta "100 metro" nobela euskaraz egon arren, bakoitzean erabiltzen diren hitzak, esateko erak eta esanahiak erabat desberdinak dira. Euskaraz dakien edonork ez lituzke hiru kasuotako maila desberdinak bereiztuko (hirurak euskaraz daude eta), baina ordenadore bidez ulertzen saiatzen badira, desberdintasunak berehala nabarituko dira. Berez ordenadoreak ez dira hizkuntzaren zailtasun guztiei aldi berean aurre emateko gauza. Emaitza probetxugarriak lortzeko, ordenadorearen lana domeinu espezifiko eta mugatu batean kokatu behar da. Etorkizunean, aplikazio mugatuzko sistemak bilduz, lor litezke ahalmen handiagoko sistema berriak, baina oraingoz martxan dauden aplikazio–motak helburu espezifikotarako dira eta beren arteko loturak hutsean daude.

1. taulan Lengoaia Naturalaren gaur eguneko aplikazio–mota posibleak biltzen dira. Bakoitzarentzat produktu erabilgarriak noraino heldu diren azaltzen da. Taulan bi aplikazio–multzo nagusi bereizten dira: alde batetik gizaki eta ordenadorearen arteko komunikazioa errazten dutenak, eta bestetik giza komunikaziorako aplikazioak. Taulako aplikazio konkretuak beste modu batera sailkatzen dira: martxan daudenak, ia erabiltzeko moduan daudenak eta oraindik proiektu–mailan daudenak.

Ondoren aplikazio–motak aztertuko ditugu, bakoitzaren deskribapena eta aplikazio zehatzak azalduz deskribatuko diren aplikazioen nazioarteko maila utzi ondoren, Lengoaia Naturalak gure inguru hurbilean lortu dituen oihartzun eta fruituak bilatuko ditugu, eta Euskara berriro plazara jalgi dadin bide bat proposatzen da bukaeran.

2. Datu–baseen galdeketa–sistemak

1. taulak erakusten duenez, dagoeneko merkatura iritsi diren aplikazioak bi motatakoak dira: datu–baseen galdeketa–sistemak (bai ordenadore erraldoietan, bai

mikroetan) eta itzulpen automatikoa. Sail honetan aztertuko dugun lehenengo mota horretan, datu–baseak erabiltzea erosoagotu egin nahi da.

Datu–base batean gai konkretu batez datu asko sartzen dira formatu zehatz baten arabera, geroago datu horien artetik zenbait baldintza betetzen dutenak erraz eta azkar atera ahal izateko. Adibidez: enpresa handi bateko enplegatuen datu–basean berehala jakin daiteke bost urtetan zeintzuek lan egin duten edo sail jakin bateko langileen zerrenda azkar lor daiteke. Baina datu–baseari galderak egin ahal izateko, bere lengoaia berezia ezagutu behar da. Betebehar hau dela eta, datu–baseen erabilpena askoz murritzagoa izaten da. Beraz oso interesgarria da galderak lengoaia naturalez egin ahal izatea, baina horrelakoetan sistemak berak galderak datu–basearen lengoaiara itzuli beharko ditu. Datu–baseei buruz azaltzen dugun guzti hau, beste edozein aplikaziotarako programez ere esan daiteke.

Alde ikaragarria dago ordenadore erraldoietarako eta mikroetarako egindako interface–en artean; bai prezioz (milioiak eta hamarnaka mila pezeta inguru hurrenez hurren) eta bai ahalmenez. Hizkuntzaren tratamendua erabat zabalago eta sakonagoa izateaz gain, sistema handietan erabiltzailearentzako laguntza eta erraztasun handiagoa dago. Erabiltzaile anitzi erantzun dakioke eta datu–base ahaltsuagoak atzitzeko aukera eskaintzen da. Mikroetan kokatutako interface–ak, guztiz desberdinak dira. Nahiz eta Lengoaia Naturalaren Proze- samenduko ikertzaile gehienek mikroetako pakete hauek gutxiesteko joera ukan, merkatuan bada zenbait sistema interesgarri. Erosleen erantzuna aztertzeko dago oraindik.

Ordenadore handietarako sortu zen lehenengo sistema, 1981. urtean Artificial Intelli- gence Corporation–ek kaleratutako "Intellect" izan zen. 1984. urteaz geroztik ordenadore- –enpresa nagusiek berori egokitu edo antzeko sistema berriak asmatu dituzte (Mathematica, IBM bere ekipoentzat, BBN eta IBS Digital–en VAX sistementzat). 1985.ean Carnegie- –Mellon–eko unibertsitateko "Carnegie Group" izeneko talde ausartak, "Language Craft" izeneko hizkuntz tresnen multzoa atera zuen. Beronek VAX sistemetarako lengoaia naturalezko interface–ak eraikitzeko aukera eskaintzen die programa–injineruei; adimen artifiziala erabiltzeko bereziki.

Mikroen munduan Lengoaia Naturalaren aitzindaria, Mexiko Berriko Excalibur Technologies enpresaren "Savvy" sistema izan zen. Bere ondorengoek bezala ez zuen hizkuntz analisirik egiten; zenbait hitz gako aurkitzean geratu egiten bait zen. Zabalkunde handirik ez zuen lortu 1984.era arte; orduan jaio bait zen IBM–PC–rako RBase paketearen "Clout" sistema. Symantec–en "Question & Answer" sistemak arrakasta ederra izan du 1986. urteaz gero. Sistema hau analisi sintaktikoa bigarren mailarako lagatzen duten "gramatika semantiko"etan oinarritzen da. Galderak oso zailak ez badira, emaitza harrigarriak lortzen ditu. Mikroetarako merkatuko azken eskaintza, Texas Instruments enpresaren "Natural Link" paketea da. Azken honetan erabiltzaileak ezin du galdetu edonola, berari aurkezten zaion menu moduko pantaila batean hitzak edo esaldi-zatiak hautatu behar ditu azaldu nahi duen galdera osatzeko. Horrela esanda, menu hutsa dela dirudi, baina bere atzetik dagoen analisatzaile linguistikoa antzeko beste sistemen mailakoa da. Pakete honen ezaugarririk onena gardentasuna da: erabiltzaileak ondo daki zeintzuk esaldi ulertuko diren eta zeintzuk ez.

3. Itzulpen Automatikoa

Itzulpen Automatikorako ere bada zenbait produktu merkatuan salgai. Horietako batek berak ere ez ditu testu literarioak itzultzen. Guztiek itzulpen teknikoan dihardute, non anbiguotasuna murritzagoa den.

Testu itzuliak beteko duen helburuaren arabera, bi multzo desberdin bereizten dira: Informazio–masa handitik edukiaren ideia orokorra ateratzen dutenak, eta zabalkuntza handia izango duten informazio zehatzak itzultzen dituztenak. Lehenengo multzorako adibidez, hizkuntza arrotzean argitaratzen den guztiaren berri ukan behar duen ikerlariak beharko lukeena dugu. Guztia ondo itzultzea denboraz edo diruz oso garestia litzateke, eta gainera zati asko ez litzaizkioke interesatuko gero. Guztiz zuzena ez den baina merkea den itzulpenaz erraz jakin ahalko luke benetan interesatzen zaion partea zein den, gero zati hori zehatz–mehatz itzultzeko. Oro har, denbora eta dirua irabaziko lirateke. Beste aldean, zabalkuntza handiko informazio zehatzen adibide gisa, etxetresna elektroniko baten erabilpenerako azalpenak ditugu. Testu horien zehaztasun eta ulergarritasuna, salgaiaren arrakastarako giltza izango dira. Beraz, kalitate handikoa izan beharko du itzulpenak.

Itzulpenaren automatizazioa ez da beti erabatekoa. Bere mailaren arabera ondoko sailkapena egiten da:

– *Erabateko itzulpen automatikoa*. Lan osoa makinaren bidez burutzen da, giza laguntzarik gabe. Errealitatea baino ametsa gehiago da gaur egun.
– *Giza laguntzaz buruturiko ordenadore bidezko itzulpena*. Lanaren arduraduna makina da, baina fase desberdinetan laguntzak eska ditzake; hitz baten adiera zuzena hautatzeko edo esaldi baten analisia nondik hasi behar den erakusteko adibidez.
– *Ordenadorez lagunduriko giza itzulpena*. Lanaren arduraduna pertsona da, baina hiztegi berezituan hitz bat edo beste bilatzeko edo testuaren formatua txukuntzeko, ordenadoreaz baliatzen da. Agian itzulpenaren zati handi bat ia laguntzarik gabe ordenadoreak egingo du, baina testua egokitzeko aurreprozesaketa edota emaitza zuzentzeko postedizioa ohizkoak izaten dira.
– *Datu–Banku Terminologikoak*. Hiztegi berezituak erabiltzeko aukera hutsa da ordenadoreak kasu honetan eskaintzen duena. Dena dela, hau ez da laguntza txikia oso testu teknikoak itzuli behar direnean, batez ere hiztegi inprimatuak baino askoz ere eguneratuago egoten direlako.

Montrealeko TAUM taldeko METEO sistema da emaitzarik arrakastatsuena lortu duena. 1977. urtean hasi zen parte meteorologikoak ingelesetik frantzesera itzultzen, testuaren %80 guztiz zuzena zelarik. Egunero oso antzekoak ziren itzulpen aspergarri hauek egiteko itzultzaileak bilatzea zaila zen, nahiz eta soldata ederrak eskaini. Urte hartatik hona lana egunero burutzen da METEOren laguntzaz. Hamaika saio egin da geroztik sistema honen diseinua beste gai batzuetara zabaltzeko, baina ezin izan da horren biribila den beste gai bat aurkitu. TAUM taldeak berak hegazkinetarako eskuliburuak itzultzeko saioak egin zituen, baina hasierako emaitza itxaropentsuek piztutako ametsak laster itzali ziren.

METAL sistema ingeles/aleman itzulpenetarako salgai dago 1985. urteaz geroztik. 1961.ean hasi ziren diseinatzen Texas–eko Unibertsitatean. Zenbait aldiz baztertu eta berrartua izan ondoren, 1980.az gero Siemens enpresa da babesle bakarra. 1986. urteaz gero Bartzelonan badago alemana/espainiera bikotera egokitzen ari den talde bat.

ALPS, Weidner eta LOGOS sistemak erabili egiten dira gaur egun; Europan batez ere. Lehenengo biak Mormoien Elizaren eraginez sortutakoak dira, bere testu sakratu ugariak errazago itzul zitezen. Hirugarrena berriz, Vietnam–eko gerran erabilitako armen eskuliburuak itzultzeko jaio zen.

SYSTRAN Institutua 1970. urteaz gero Itzulpen Automatikorako tresna–saltzaile nagusia izan da. NASA, Europako Ekonomi Elkartea, General Motors eta Xerox dira bere bezerorik ezagunenak. Europako Ekonomi Elkarteak egokitzapen neketsua behar izan zuen (100.000 hitzeko hiztegia definitu behar bait zuen) frantzes/ingeles itzulpena ahalbideratzeko. Egun 20 itzultzailek erabiltzen dute sistema hau Luxemburg–en, hilabetean milaren bat orrialde ingeles/frantses, frantses/ingeles eta ingeles/italiera bikoteetarako itzultzen dutelarik. Kanadako General Motors–ek eskuliburuak itzultzen ditu ingelesetik frantsesera. 130.000 hitzeko hiztegia definitu ondoren, itzultzaileen lana 3 edo 4 aldiz arinagoa zen, eguneko 1000 hitzeko mailaraino helduz. SYSTRANen oinarri informatikoa, guztiz atzeratua dago; 1960.eko hamarkadako teknologia erabiltzen bait du.

Europako Ekonomi Elkarteak, SYSTRAN sistema bere itzulpen–beharrak betetzeko tresna aski ahaltsu ez zela ikusita, EUROTRA proiektu berria abiarazi zuen 1978.ean. Europako 9 hizkuntza nagusiak hartuta (euskara, katalana eta antzeko hizkuntzak ez daude) edozeinetatik beste edozeinetara itzultzeko gauza izango zen, espezifikazioaren arabera. Hasieran finkatu ziren epeak ez dira bete hizkuntza guztietarako (frantzesa, ingelesa, alemana eta danierarako soilik). Hizkuntzalari eta informatikarien arteko proportzioa guztiz desorekatuta dago lehenengoen alde. Arrazoi hauengatik, zenbait behatzaileren ustez proiektu honetatik ezin daiteke, epe laburrerako behintzat, fruitu zehatzik espero.

Grenobleko GETA taldeak 1961. urteaz gero dihardu eremu honetan. Hasieran errusieratik frantzeserako itzulpenak burutu ziren. Gero ingelesa, alemana eta arabiera edo malaysiera bezalako beste hizkuntza batzuk ere aztertu dira. Azken urte hauetan, Frantziako gobernuaren diru–laguntzak direla medio adimen artifizialeko lengoaia eta tresnekin sistema birmoldatzen ari dira.

Itzulpen Automatikoak garrantzizko papera du Japoniako bostgarren belaunaldiaren proiektuan. Bertan adimen artifizialak eta itzulpen automatikoak elkarrekin lan egiten dute. 1985.erako trilioi bat yen gastatuta zegoen ikergai honetan. Orduan Europa eta Estatu Batuetan 12 talde ikertzaile biltzen ziren iharduera honetan eta aldi berean Japonian 18 ziren. Fujitsu, Hitachi eta Toshiba erraldoiak bereziki interesatuta daude eskuliburuen itzulpenaz, eta beren taldeak osatu dituzte.

4. Testuen eduki–araketa

Aplikazio–mota honetan, testuak barruan daukan datu bat bilatzeko edo laburpen bat lor- tzeko aztertzen da. Oraindik ez dago merkatuan era honetako sistemarik salgai, baina agertzear daude batzuk. Cognitive Systems enpresaren Atrans paketeak bankutako telex–en informa- zioak irakurtzen ditu. Antzeko sistema bat garatzen du Carnegie–Mellon–eko taldeak. Cogni- tive Systems–ek Estatu Batuetako Kostazaintzarako egiten duen sisteman, untziei buruzko mezuak hartu eta munduko itxasuntziei buruzko datu–basea eguneratzeko erabiltzen du.

5. Testu–edizioa

IBM, Macintosh eta bestelako PC arruntetan ingelesa, frantsesa, espainiera eta beste hizkuntza nagusientzako ortografi zuzentzaileek bete dituzte urte batzuk merkatuan. Laborategietan bukatzear dauden pakete berriek, idazkera– eta sintasi–erroreak ere zuzenduko dituztela dirudi. Nahiz eta errore guztiak harrapatu ez, laguntza ederra eskaintzen dio eskutitzak edo bestelako txostenak idatzi ohi dituenari.

6. Elkarrizketarako intefaceak

Aplikazio–mota honetako sistemek, ordenadore eta gizakiaren arteko komunikazio erosoa ahalbideratzen dute. Galdera eta erantzunez osatutako elkarrizketa ulertu ahal izateko, partaideen planak eta helburuak aztertzeko tresnak beharrezkoak dira. Hiztun bakoitzak momentu bakoitzean zer dakien eta zer nahi duen zehaztu behar da eta gainera ezagumendu horiek dinamikoki eguneratu behar dira elkarrizketa aurrera joan ahala.

Honelako sistemak inplementatzen zailak dira. Helburu orokorrekorik ez da salgai egongo urte luzetan, baina badira aplikazio zehatzei lotuta dauden batzuk. Gehienak adimen artifizialeko erabilpenei lotzen zaizkie.

Ordenadorez Lagunduriko Irakaskuntzarako SOPHIE sistemak, gaizki dabiltzan zirkuitu elektronikoak diagnostikatzen laguntzen dio ikasleari. Elkarrizketa osoa lengoaia naturalez egiten da.

MYCIN sistema adituak, elkarrizketa baten bidez lortzen ditu diagnostikatuko duen gaixotasunaren sintomak.

7. Ahozko idazmakina

Azken aplikazio–mota honek zailena dirudi, baina lortuz gero aldaketa ikaragarriak sortuko lituzke. Zailena, hizkuntza idatzia ulertzeko arazoei ahozko hizkuntzaren anbiguetateak eransten zaizkiolako da: hitzak ez dira guztiz bereizten hitz egiterakoan, esaldietako hasiera eta bukaera erdikoa baino intentsitate txikiagoz ematen dira eta gainera seinale fisikoen zaratak ohizko oztopoak izaten dira.

Ikertalde ospetsuenek (IBM, Carnegie–Mellon, MIT, ...) saio neketsutan dihardute *ahozko idazmakina* lortzeko, baina beren lorpena oraindik ez dago gertu. Gaur egun uler daitezke makinaz persona zehatz batek emandako hitzak (esaldiak ere bai noizbait), baina beste pertsona batenak ulertu ahal izateko berriro hezi behar da makina, hiztun berriak aurrez prestatutako testu bat irakurriz eta hitzak espreski bereiztuz. Onartzen den hiztegia, mugatua da. Transkripzio–erroreen kopurua handia izaten da.

8. Euskal Herriko egoera

Aurretik azaldutako aplikazio guztiak ingeleserako eginak dira. Salbuespen batzuk badira; alemaniera, frantzesa, japoniera eta daniera lantzen dituztenak batez ere. Espainiera ere hizkuntza interesgarri bihurtu da azkenaldi honetan; herri aurreratuetako produktuentzat espainieradun erostunen kopurua handia delako batez ere. Egun Bartzelonan itzulpen automatikoari buruz hiru proiektu zabaldu dira, beste proiektu erraldoien sukurtsal moduan: METAL (Siemens–ena alemaniera/ingelesa itzulpenak burutzen dituena), EUROTRA (Europako Ekonomi Elkartearena) eta FUJITSU

japoniarra. Madrileko IBM–k MENTOR proiektuan ingeles, espainera eta hebraierarako sistema bat garatzen du. Frantsesa egoera hobean dago; ingeles/frantses, alemaniera/frantsez eta errusiera/frantses bikoteak aztertuak izan bait dira Kanadako TAUM–METEO sisteman, Grenobleko GETAn eta EUROTRA barruan. Talde hauen esperientzia Espainiakoena baino askoz ere luzeago eta sakonagoa da.

Katalanerako oraindik ez dago proiektu sakonik, baina analisi morfologikoa guztiz inplementatuta dute, ordenadore bidez automatikoki egin ahal izateko. Bestetik Bartzelonan kokatu diren itzulpen automatikorako hiru proiektuen itzalpean katalanerako tresna berriak garatu nahi lituzkete bertako langile katalanzale batzuk.

Oraingo hamarkada hau hasi zen arte, Euskal Herrian ez zen ezer entzun ikerkuntzaren eremu honetaz. Ordutik hona burutu diren lanak ondorengoak dira:

– UZEIko Euskal Term. Datu–base terminologikoa, non UZEIren hiztegi berezitu guztiak erraz kontsulta daitezkeen.
– Joseba Abaitua hizkuntzalariak Manchester–en aurkeztu zuen tesia. Euskararen morfologia eta sintaxirako gramatika lexiko–funtzional bat proposatzen du. Egun Bartzelonako FUJITSUren itzulpen automatikorako proiektuan ari da lanean.
– Donostiako Informatika–Fakultatean 7 partaideko talde bat ari da bide berri hauek urratu nahian. Hasiera CAPRA proiektuaren eskutik etorri zen. Proiektu horretan, ordenadore bidez ordenadorearentzako programak idazten irakatsi nahi da. Bere barruan bi doktorego–tesi gorpuzten dira. Batean problemen enuntziatuak lengoaia naturalez automatikoki ulertzeko sistema bat eraiki zen eta bestean ikasle eta ordenadore–tutorearen arteko komunikazioa lengoaia naturalez burutzen da. Sistema gaztelaniarako egin da, baina euskarari ere egokitu zaio zenbait modulu. Aurten proiektu berria abiarazi da UZEI Institutuarekin eta APIKA informatika–enpresarekin batera. Proiektu honen helburuak bi dira: euskaraz idatzitako testuentzako zuzentzaile ortografikoa eta euskararako analisatzaile morfologiko orokorra burutzea.

9. Euskararen etorkizunerako bidea

Datu baseen galdeketa–sistemak, ordenadorez lagunduriko irakaskuntz sistemak edo itzulpen automatikoa euskaraz ikusi ahal izateko, azpiegitura guztia egiteko daukagu.

Informatikariak eta hizkuntzalariak trebatu beharko dira arlo honetan, gero taldelanean eta aplikazioen artean dagoen lehentasunari jarraituz azken helburuetara iritsi ahal izateko.

Oinarri–oinarrizko tresnak honako hauek lirateke: hiztegi informatizatuak, zuzentzaile ortografikoa, analisatzaile morfologiko automatikoa eta analisataile sintaktikoa. Hizkuntz tratamendu errazetarako sistemak eraikitzea litzateke bigarren pausoa. Testu luzeak ez direnerako, esaldi–mailako elkarrizketak ulertu eta gidatuko dituzten sistemak asma daitezke, gai mugatu baten barruan eta zenbait muga linguistikorekin. Aplikazio ugari aurki liteke administrazioan muga hauen barruan (datu–baseen galdera–erantzun moduko sistemen bidez batez ere), baina baita Ordenadorez Lagunduriko Irakaskuntza edo Sistema Adituen bidez ere. Bigarren pauso paraleloa, Itzulpen Automatikoa litzateke.

Erreferentziak:

- "A survey of Machine Translation: its history, current status and future prospects"
  J. Slocum. Computational Linguistics Janaury-march 1985

- "Natural Language Understanding"
  J. Allen. The Benjamin Cummings Publishing Company. 1987.

- "Natural Language Computing: the commercial aplications"
  T. Johnson. Ovum Ltd. London 1985.