Linguistika eta matematika

Txillardegi

Gero eta maizago gertatzen da gaur, linguistikari buruzko liburu bat eskuetan hartu, eta... ekuazioak, →matrize erako ikur-zutabeak, multzo-ebaketak, ulerkuntza zaila duten eskemak eta irudiak, eta abar, aurkitzea.

Hizkuntz jakintzak Giza-Zientzien aintzindaritza beregana bildua duenez geroztik, zaildu egin bide da; eta zorroztu ala, matematikorrago ere agertzen zaigu.

Linguistikaren sail edo adar batzuk bultzatzen dute batera alderdi horretara:

1. →Linguistikaren alorrean *Estatistikak* eragin dituen ikerketak daude batetik: →hitzen edo →fonemen →maiztasunak (→Zipf-en legeak bereziki), idatz-moldeen zenbatasunezko azterketak, eta abar. Azterketa horiek ondorio garrantzitsuak izan dituzte maila batzutan: →hizkuntzen irakasbideen prestaeran, hizkuntza barruko →koerlazio-bilaketan, eta abar.

2. Badago ere, bigarrenik, →mezuen igorpenak berak *Telekomunikazio-Teknikan* sortu duen Informazio-Teoriaren eragina. Markov-en kateek eta Shannon-en teoremek, horretara, beren matematika-kutsua utzi dute Linguistika berrian: «→entropia», «→bit unitatea», «→erredundantzia», logaritmo bitarren beharra, eta abar.

3. Badago, hirugarrenik, nahiz bere funtsean kritikaturik ere, →*Glotokronologia* deritzon ekarpena. Swadesh-en teoriaren arabera, eta oinarrizko hiztegiaren kantitate-azterketa eginda, jakin daiteke jatorri bereko bi hizkuntza desberdin (greziera eta armeniera, eman dezagun) noiz «berezi» ziren.

4. →*Multzoen Teoria* izan da laugarren bultzatzaile bat. →Zenbatasunezko azterketak hizkuntzaren alorrean askotan halako desegokitasuna baldin badu ere, egokiago bide datozkio nolakotasunezko azterketak. Hots, Matematika berriak, Multzoen Teoriak, Logika matematikoak,

eta Logika soilak (Wittgenstein, B. Russell), tresna, izen eta →algoritmo zorrotzak eskaini dizkiote gaurko Linguistikari. →Fonologia bera, beste alde batetik, oso dotoreki azal daiteke Multzo-Teoriaren arabera.

5. Eta bostgarren eragilea, nabarmenena agian, →Chomsky bera da, eta *Ameriketako Linguistika berri* osoa ere. Perpausaren antolakera eta →hiztunaren sormena matematikoki azaldu nahiz, zeharo eman dio Generatibismoak Linguistika berriari bere Logika, Informatika eta Matematika kutsu nabarmena: «teorema», «→axioma», eta horrelako hitzak, topatzen dira nonnahi; baita →zuhaitz sintagmatikoak han eta hemen, eta matrizeak, eta abar.

Orrialde gutxitan mundu oso horren berri ematen zail bada ere, joeren lerro nagusi batzu bederen agerrerazten saiatuko gara; baita, gai bakoitzari dagokionez, bibliografia apur bat ematen ere.

I. GLOTOKRONOLOGIA

Eskualde eta mintzaira-mota guztietan aurkitzen dira hizkuntz →maileguak. Gertakari orokor honen arabera, →hitzak (edo-ta →fonemak) esportatu egiten dira, herritik herrira mailegatu; eta bazeuden hizkuntzatik, ez zeuden hizkuntza batetara hedatzen: ingelesezko «meeting» delakoa, «mitin» itxura hartuz, espainierara igaro da; «taladro» erdal hitza, «daratulu» itxura hartuz, euskarara sartu da; eta abar.

Jakina denez, hizkuntz elementuak eremu egituratuago batetara joan behar, eta nekezago barneratzen; eta, alderantziz, hizkuntz egituretatik bere buruz eta bereziago gertatu behar, eta aisago ere onar eta barneratzen. Fonemen →sistema, horretara, askoz ere →iraunkorrago da, tresneriaren hiztegia baino; eta askoz ere aisago gertatzen dira hitzen maileguak, →morfemen edo fonemen maileguak baino.

Hots, azterketa batzu egin ondoren, hipotesi honetara heldu zen Morris Swadesh hizkuntzalaria: «oinarrizko hiztegia» ongi hautatzen baldin bada, politika eta kondaira-baldintzapenak eta eraginak nolanahikoak izanik ere, kanpotikako hitz-mailegutza abaila edo *erritmo berean* egiten dela hizkuntza guztietan. Mendeoroz, hitz batez, eta Swadeshen iritziz, hitz-kopuru *berbera* aldatzen dute hizkuntza guztiek (beti ere «oinarrizko hiztegi» horri gagozkiolarik).

Hipotesi hori egia baldin bada, hortaz, piska bat «carbono C-14» aren bitartez →datazio eta kronologi bilaketan egiten den bezalatsu jokatuz, hizkuntzen bilakaera-gorabeherak ere aztertu ahal izango ditugu.

Legea, beraz, berretazkoa da; eta honela idatz daiteke:

$$c = r^t$$

Edo, logaritmoak hartuz gero:

$$\log c = t \cdot \log r; \text{ eta;}$$

$$t = \frac{\log c}{\log r}$$

t: alderatzen diren hizkuntza (berbera)ren bi aldi edo fase horien artean joandako denbora.
c: hizkuntza berberaren bi mintzaira-aldiek atxiki duten partea.
r: denboraldi-unitatean oinarrizko hiztegi horretan, jatorrizko hitzaz atxikitzen dena.

Zehatz ditzagun orain oinarrizko hiztegi hori, eta formulan agertzen diren koefizienteak.

A) OINARRIZKO HIZTEGIA

Glotokronologiaren ikerketak sakondu ala, zorroztu egin ditu Swadesh-ek bere koefizienteak eta hitz-zerrendak.

Oinarrizko hiztegitzat, horrela, hiru hitz-zerrenda desberdin proposatu izan dira: lehenengoak 215 ditu, bigarrenak 200 hitz, eta hirugarrenak 100 hitz. Hona hemen ehun hitzetako zerrenda:

— guzti, hauts, zuhaitz-azal, sabel, handi, txori, hausiki, beltz, odol, hezur;
— erre, hodei, hotz, etorri, hil, zakur, edan, lehor, belarri, lur;
— jan, arraultza, begi, koipe, luma, su, arrain, hegan, oin, eman;
— on, berde, ile, hazi (= bihi), eseri, larru, belaun, ilargi, borobil, esan;
— esku, buru, entzun, bihotz, ni, erhan (hil, nork), jakin, hosto, etzan, gibel;
— luze, zorri, gizon (= gizakume), asko, haragi, mendi, aho, izen, lepo, berri;
— sudur, ez, bat, lagun, euri, gorri, bide, sustrai, hondar, ikusi;
— lo-egin, txiki, ke, zutik (egon), izar, harri, eguzki, igeri, buztan, hori (hau/hori);
— hau, hi, hiru, mihi, hortz, zuhaitz, bi, ibili, bero, ur;
— gu, zer, zuri, nor, emakume, hori (kolore), altzo, erpe, bete, adar.

Eman dezagun orain *H hizkuntza bakar baten* bi hizkuntzaldi alderatzen ditugula:

H_1 (t_1 urteko hizkuntza-molde zaharra)
H_2 (t_2 urteko hizkuntza-molde berria)

Eman dezagun ere 100 hitzen zerrenda erabiltzen dugula; eta azterketa sakon bat eginez gero, hau aurkitzen dugula: H_1 hizkuntzaldian genituen ehun jatorrizko hitz horietatik, *c* hitz «jator» gelditu direla hizkuntza berrituan; eta, hortaz, *100 — c* hitz aldatu direla: mailegu bidez, kanpotik sartu dira *100 — c* hitz arrotz, bertako zaharrak aihenatuz.

Hortaz, hizkuntza berberaren bi hizkuntzaldi edo molde horien artean, $(t_2 - t_1)$ urte pasa ondoren, hau dugu:

$$Sw\ 1 — \qquad t_2 - t_1 = \frac{\log c}{\log r}$$

c ezagutzen dugu: jatorri bereko hitz jatorren kopurua.

Zenbat balio du *r* horrek? Hauxe da Swadeshek, anitz azterketaren ondoren, ezagutarazten diguna:

$$r = 0{,}854 \pm 0{,}004 \text{ (/ mila urteko)}$$

Swadeshen legea, beraz, honela idatz daiteke;

$$t_2 - t_1 = \frac{\log c}{\log 0{,}854}$$

Logaritmo hori, noski, edozein oinarritakoa izan daiteke; baina bietan oinarri berekoa.

$t_2 - t_1$: bi hizkuntzaldi horien arteko denbora, *r* neurtzeko hartutako unitate berean; kasu honetan, beraz, mila urtetan.

1. Argi dezagun hau *etsenplu* baten bitartez.

Eman dezagun H_1-eko hiztegiaren arabera, % 30 galdurik ikusten dugula H_2 hizkuntzan. Hortaz:

$$c = \%\ 70 = 0{,}7$$

$$t_2 - t_1 = \frac{\log 0{,}7}{\log 0{,}854} = \frac{-0{,}1549}{-0{,}0685} = 2{,}261 \text{ (mila urte)}$$

Beraz, $t_2 - t_1 = 2.261$ urte.

r koefizienteaz gorabeherak badaude, jakina. Hona hemen hizkuntza ezagun batzuren arteko azterketak ematen duena:

kopto/egipziera:	0,76
latinera/errumaniera:	0,77
latinera/frantsesa:	0,79
latinera/portugesa:	0,82
latinera/italiera:	0,85

Ulertzen errazak diren arrazoinengatik, areago atxiki du italierak jatorrizko latina, esate baterako, errumanierak baino.

Hori dela eta $r = 0{,}81$ hartzen dute batzuk (nahiz 100 hitzen zerrendari koefiziente handiagoa egokitu).

2. Lehengo etsenpluan, $r = 0{,}81$ hartuz gero, hau dugu:

$$t_2 - t_1 = \frac{-0{,}1549}{-0{,}0915} = 1{,}629 \text{ (mila urte)}$$

$$t_2 - t_1 = 1.629 \text{ urte}$$

Swadeshen legeak, hitz batez, *gutxi gora-beherako* ondorioak besterik ezin eman ditzake. Eta gisa da. Oinarri duen hipotesia ezin bait da xeheki eta hertsiki hartu.

B) SWADESHEN LEGEA

Swadeshen legea, era berean, *jatorri bereko bi hizkuntza* bereziren arteko berezkunea kalkulatzeko ere erabil daiteke.

Berezkuneko t_0 izan bazen, eta orain t_1 urtea baldin bada, hizkuntz adar bakoitzetik joanez gero, hau izango dugu:

$$c_1 = r(t_1 - t_0)$$
$$c_2 = r(t_1 - t_0)$$

eta biok orain alderatuz gero:

$$c_{12} = c_1 \cdot c_2 = r^2(t_1 - t_0)$$

edo, logaritmoak hartuz:

$$Sw\ 2 - \qquad t_1 - t_0 = \frac{\log c_{12}}{2.\ \log r}$$

Argi dezagun hau beste etsenplu batez.

Eman dezagun bi hizkuntza jatorkide alderatuz, eta oinarrizko hiztegiari buruz beti, % 63 hitz jatorkide edo «cognate» aurkitzen dugula. Esate baterako: esp. verde, fr. ouvrir (lat. viridem); esp. lleno, fr. plein (lat. plenum), hitz «cognate» dira.

$$t_1 - t_0 = \frac{\log 0{,}63}{2.\ \log 0{,}854} = \frac{-0{,}2007}{2.(-0{,}0685)} = 1{,}465 \text{ m. u.}$$

$$t_1 - t_0 = 1.465 \text{ urte}$$

Hitz «cognate» edo jatorkideen zerrenda ondo zehaztuz gero, ondorio egokiak ematen bide ditu Swadeshen ikerpideak; batez ere 500 eta 2.500 urteren artean.

II. ZIPF-EN LEGEA

Hiztegien estatistika-lege bat da Zipf-ena.

Baina, agerreraz daitekeenez, Informazio-Teoria oinarritzat hartuz, Zipf-ena ondorio gisa froga daiteke.

Egia esan, badaude izen horretxetaz eman ohi diren beste estatistika-lege batzu; eta ez bakar bat. Eskuarki, ordea, hitzen *maiztasuna* bera, eta maiztasunen araberako lerroko *gradua* lotzen dituen legeari deritzo «*Zipf-en Legea*».

A) Eman dezagun edozein hizkuntzatako «→corpus» bat: liburu bat, esate baterako. Azter dezagun lehendabizikorik hitz bakoitza *zenbat aldiz* agertua den textu horretan; eta hitz bakoitzari datxekion zifra berezi horrek hitzaren *maiztasuna* neurtuko du (*f*). «Etxe» hitza, eman dezagun, 326 aldiz agertu baldin bada, 326 izango da «etxe» hitzaren maiztasuna.

Har ditzagun orain liburuan agertutako hitz horietxek; eta idatz ditzagun, tajuturik, zerrenda luze batetan. Lehenengo hitza, maiztasunik handiena erakutsi duena izango da. Bigarrena, bide beretik, lehenengoa idatzi ondoren, maiztasunetan nagusi agertzen dena izango da. Eta abar. *Maiztasunen zerrenda* prestatuko dugu horrela, eta hitz bakoitzari bere *gradua* (*r*) erantsiko zaio. Zerrendan zazpigarren agertzen den hitzaz, 7 gradukoa dela esango dugu. Eta «etxe» hitz hura 174. lerroan agertu baldin bazaigu, 174 gradukoa dela esango dugu.

Bide enpirikotan barrena, hasieran, eta saiaketak erakutsi zuenez, «corpus» guztietan betetzen da ondoko ekuazio hau:

$$r \cdot f = k \text{ (konstante)}$$

1916-az geroztik jabetu zen lege honetaz estenografiaz ari zen Jean Estoup. Matematika-lege gisa, ordea, G. K. Zipf hizkuntzalariak eman zuen lehendabizikorik.

Eman dezagun Joyce-ren «Ulysses» nobela famatua. Bertan agertzen diren hitzen zenbaketa eta zerrendaketa egiten baldin badira, hau aurkitzen da:

r (gradua)	*f* (maiztasuna)	*r.f* (biderkaketa)
10	2.653	26.530
20	1.311	26.220
100	265	26.500
300	84	25.200
1.000	26	26.000
5.000	5	25.000

Kasu honetan, beraz: $r \cdot f = 26.000$.

(Estatistika-bideak baliatuz gero, jakina denez, hitz baten edo besteren irregulartasuna behar bezala hartzeko, on izan daiteke zerrendako hitzak hamarnaka edo neurtzea; eta ez banaka).

Zerrendako lehenengo graduetako hitzak baztertuz gero (gehien agertzen direnak, beraz), ongi samar betetzen da legea; eta edozein izkributan aurki daiteke gauza bera.

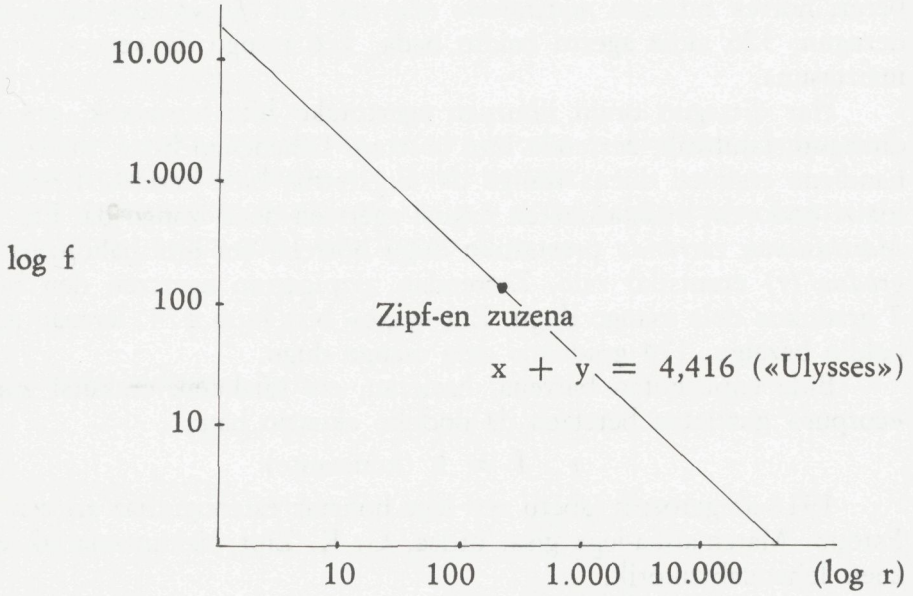
Logaritmoak hartuz gero, ordea, hau dugu:

$$\log r + \log f = k$$

Eta grafiko batetan eskala logaritmikoak erabiliz gero:

$$x = \log r$$
$$y = \log f; \text{ eta, beraz: } x + y = K$$

Logaritmotan eta grafikoki emanez gero, hortaz, zuzen baten itxuran agertzen da Zipf-en Legea:



(Marrazki horretan agertzen den zuzenean: K = log 26.000 = 4,416).

Maiztasun handietan gertatzen diren desarautasunak matematikoki bildu nahiz, formula zailago bat proposatu du Mandelbrot-ek:

$$(r + b)^a \cdot f = k \quad (a > 1)$$

Baina Informazio-Teoriak Zipf-en Legera eramaten du (ikus «Le Langage», Pléiade, 46/56 eta 145/157).

B) Eman dezagun orain textu hau:

L = hitzen kopuru osoa
n = hitz desberdinen kopurua.
Jakina = $n \leq L$

Batetik: $f_1 + f_2 + f_3 + \ldots + f_n = L$
eta, bestetik: $f_1 \cdot 1 = f_2 \cdot 2 = f_3 \cdot 3 = \ldots = f_n \cdot n$
Beraz: $f_n \cdot n = K$
($K = 26.000$ aurreko etsenpluan)

Edo-ta:

$$\frac{K}{1} + \frac{K}{2} + \frac{K}{3} + \ldots + \frac{K}{n} = L$$

$$K = \frac{L}{1 + \frac{1}{2} + \frac{1}{3} \ldots + \frac{1}{n}}$$

Edozein textutan aurkitzen diren *maiztasunak* sorta armonikoaren araberako balioak dira.

1. Lehengo adibidean: $K = 26.000$
Beraz, kalkula dezagun f_{10}, esate baterako:

$$f_{10} \frac{26.000}{10} = 2.600$$

Hots, taula enpirikoan hau dugu:

$$f_{10} = 2.653$$

Ia-ia Zipf-ek proposa lezakeen kopurua.

III. INFORMAZIO-TEORIA

Eman dezagun *komunikabide* bat, →mezuen igorbide bat: Morse-sistema, eman dezagun. Batzutan marra (—) erabiliko dugu, beste batzutan berriz puntua (.); eta, biok nahasiz eta herrenkatuz, *letrak* (A = .—, eta abar), eta *mezuak* ere bidaliko ditugu: hariz, irratiz, edo bestetara.

Era berean jokatzen dugu edozein seinale-→sistematan. Eta gauza bera egiten dugu ere, noski, hitzegiten dugunean. Lehendabizi eman dezagun, «z» fonema ebakitzen dugu; gero «u» fonema; gero «b»; eta gero, azkenik, «i». Lau fonema horien ilarak «zubi» ematen du.

A) Eman dezagun orain (haurtzaroko joko hartan bezalaxe) «z» letra eman digutela. Zer izan ote daiteke segida? Zein fonema ote dator

ondotik? Beharbada «zaku» esan nahi digute; edo «zapi», edo «zatoz», edo-ta «zoaz», edo abar. «z» fonema horrek *informazio* txikia ematen digu.

Eman dezagun orain *u* erantsi, eta «zu» asmatu dugula. Zer ote da segida? Ez ote da «zure»? Ez ote da «zume», «zuhaitz» edo «zulo»? Ez ote «zubi»? Asmatzeko probabilitatea txikia da oraindik; baina lehen baino *informazio* handiagoa badugu. Orain badakigu, esate baterako, «zaku», «zapi», «zerri», «ziri» eta «zori» ez direla posible.

Eman dezagun orain «zub» mezu-→egoeran gaudela. Zer ote datorke atzetik?, «zubr» ez da posible, «zubt» ere ez, «zubo» ez da ezer. Ia-ia segurutzat jo daiteke «zubi» dela. Laugarren fonema honek, hortaz, *ez digu* informazio handirik eransten. Ia-ia *segurutzat* eman daiteke; eta, horregatixek, ia-ia *hutsa* da dakarkigun informazioa. *Seguru denak ez du batere informatzen. Espero ez genuenak, aldiz, erruz informatzen du.*

Informazio-Teoriaren kakoa horretantxe datza, hain zuzen:

$$I_a = \log_2 \frac{1}{p_a}$$

a seinalearen probabilitatea p_a baldin bada, *a* horrek dakarren informazioa (I_a) probabilitate horren alderantzizkoaren logaritmo →bitarra da.

1. Zergatik probabilitatearen *logaritmoa,* eta ez alderantzizkoa bera? Segidan agertuko denez, erosotasunari dagozkion arrazoinengatik bakarrik; probabilitateen metaketa (mezuetan bezala) *biderkaketaz* neurtzen bait da; eta ez batuketaz. Arrazoin praktikoak ere, logaritmoetan bitarrak hain zuzen hautatzeko.

2. Kasurik *xinpleena,* noski, txanpon batena izan daiteke; edo, eletrika-tresnerian bezala, elektrika pasa bai (+), eta eletrika pasa ez (—), (eta biak probabilitate berekoak izanik).

Kasu horretan, gertaera bakoitzaren probabilitatea 1/2 izanik, hauxe da egoera:

$$I_a = I_b = \log_2 \frac{1}{p} = \log_2 \frac{1}{1/2} = \log_2 2 = 1$$

Aukera-behar horretan, beraz, $I_a = I_b = 1$. Informazio-unitatea, horretara, «→bit» izenaz ezagutua, horrela definitzen da. Eta aukera xinple horretan ($p_a = p_b = 0{,}5$) hau dugu:

$$I_a = I_b = 1 \textit{ bit}$$

3. Hortaz, eta edozein sistematan, *a* seinalea agertzeko probabilitatea, eman dezagun, 0,125 baldin bada, zer *informazio-kopuru* erantsi du *a* seinale horrek?

$$I_a = \log_2 \frac{1}{0{,}125} = -\log_2 0{,}125 = \log_2 2^3 = 3 \textit{ bit}$$

eta, orokorki: $I_a = \log_2 \frac{1}{p_a} = -\log_2 p_a$

4. Logaritmo-taulek eta esku-kalkulagailuek, jakina denez, Briggs-en logaritmoak eman ohi dituzte eskuarki; alegia, hamartarrak, eta ez bitarrak. Hots, erraza da batetik bestera igarotzea. Aski bait da ekuazio hau kontutan hartzea:

$$\log A = \log 2 \,.\, \log_2 A; \text{ eta, } \log_2 A = \frac{\log A}{0{,}301}$$

5. *Adibidez:* Eman dezagun hiru seinale *(a, b, c)* dituen mezu-sistema bat; eta seinaleen probabilitate-taula ondoko hau dela:

$$p_a = 1/2$$
$$p_b = 1/3$$
$$p_c = 1/6$$

(Noski: kasu honetan, eta denetan, $p_a + p_b + p_c = 1$)
Agertua delarik, zer informazio-era eransten dio mezuari *a* seinaleak?

$$I_a = -\log_2 0{,}500 = \frac{0{,}301}{0{,}301} = 1{,}00 \text{ bit}$$

Eta, bide beretik:

$$I_b = -\log_2 0{,}333 = \frac{0{,}476}{0{,}301} = 1{,}58 \text{ bit}$$

$$I_c = -\log_2 0{,}167 = \frac{0{,}776}{0{,}301} = 2{,}58 \text{ bit}$$

(Oharra: espero zitekeenez, informazio-kopuru handiagoa dakar *c* seinaleak, *a* seinaleak baino).

Hizkuntza batetan, hortaz, eta lehenengo hurbilketan, aski izango da fonemen probabilitate-taula ezagutzea (p_a, p_b ... p_z), fonema bakoitzak, agertzean, eransten duen informazio-era jakiteko.

B) Entropia. Seinale hau edo hura mezuaren katean agertu denean, ordea, beste guztiak ezin ager.

1. Lehenengo adibidera bihurtuz, «zu-» mezu-→egoeran ginenean, eta *b* agertzean, ez da *f* agertu, ez *u*, ez *p*, eta ez gainerakoak. Kolpe edo mezu-seinale *bakoitzeko*, hortaz, seinale horretxen probabilitatea zegoen berori agertzeko.

Zeini bere pisua emanez beraz, hiru seinale dituen sistema horretan hauxe izango da «→entropia» →sinbolo *bateko:*

$$H_1 = -(p_a \cdot \log_2 p_a + p_b \cdot \log_2 p_b + p_c \cdot \log_2 p_c)$$

Eta, orokortuz:

$$H_1 = -\sum (p_i \cdot \log_2 p_i) \text{ [bit / seinale]}$$

2. *Adibidez:* Lehengora berriro itzuliz:

$p_a = 0{,}500$; $p_b = 0{,}333$; $p_c = 0{,}167$

Zein da, *seinale bakoitzeko,* sistema horrek duen entropia?

Presta dezagun taula hau:

	p_i	$-\log_2 p_i$	$-p_i \cdot \log_2 p_i$
a	0,500	1,00	0,500
b	0,333	1,58	0,527
c	0,167	2,58	0,431
			1,458 [bit/sein.]

Kontuz! Mezu batzutan, *b* eta *c asko* denean, informazio-kopuru *handiagoa* gertatuko da batez-bestean; eta bestetan, berriz, *a* asko denean, *txikiagoa.* Mezu guztien erdigunea 1,458 da, ordea.

3. *Entropiaren maximoa.* Bi seinale bakarrik daudenean *(a, b),* biok probabilitate berekoak izatean gertatzen da sistemaren hoberena:

$$I_{max} = -2 \cdot 0{,}500 \cdot \log_2 1/2 = 1{,}000 \text{ [bit/sein.]}$$

Eta lehengo adibidera bihurtuz, hiru seinaleei probabilitate berbera emanez gero: $p_a = p_b = p_c = 1/3$; hau dugu:

$$H_0 = 3 \cdot 0{,}333 \cdot 1{,}58 = 1{,}58 \text{ [bit / sein.]}$$

Hauxe da edukieraren *maximoa hiru seinalez.*

Eta *n* seinalez hauxe da maximoa: $I_{max} = \log_2 n = H_0$

Informazio hutsaren aldetik, seinale guztiei probabilitate berbera ($p = 1/n$) legokiekeen hizkuntza litzateke onurakorrena.

27 fonema dituen hizkuntza batetan, esate baterako, hauxe litzateke informazio-edukierarik «onena»:

$$H_0 = \log_2 27 = 4{,}76 \text{ [bit / fonema]}$$

Praktikan 4,00-ren inguruan kokatu ohi da edukiera hori. Alde batetik fonema batzu besteak baino maizago erabiltzen direlako; eta bestetik dagoen «→erredundantzia»rengatik; edozein fonema ezin bait daiteke edozein fonemaren ondotik agertu.

H_0 kopuruari (*p* guztiek 1 balio dutenean),

$$\frac{}{n}$$

«*zero graduko entropia*» deritzo.

H_1 kopuruari, berriz (p desberdinak), «*bat-graduko entropia*» deritzo.

Gorago ere esan denez:

$$H_1 \leqslant H_0$$

Era berean, H_2, H_3, eta abar, definitzen dira. Eta beti:

$$H_0 \leqslant H_1 \leqslant H_2 \leqslant H_3 \ldots$$

IV. GRAMATIKA SORTZAILEAK

Beste leku batzutan aztertzen dira berezikiago bai →Chomsky bai «→morfosintaxia». Hemen ere aipatu behar dira gramatika «sortzaileak»; baita →hiztunaren sormena matematikoki azaltzekotan hartu den bidea.

Nola ote liteke haur batek sekula entzun ez duen perpaus bat xuxenki asmatu eta eratzea? Skinner behavioristaren aurka, eta ordura arte emandako azalpenen kontra, →Chomsky-k hau dio: hiztun horrek, barneraturik, nahiz zeharbidez eta ohargabeki bada ere, ikasia duen hizkuntza horren GRAMATIKA SORTZAILEA darama; eta honegatixe asmatzen ditu sekula ez esandakoak, eta ongi asmatzen.

Gramatikak, hitz batez, perpausen MOLDABIDEA azalduko du, sortzailea izango da; eta hau ez deno, taxonomia hutsa eta antzua da, ez bait du mintzamena bera azaltzen.

A) GRAMATIKA SORTZAILEAREN OSAKINAK

Matematika Logikaren arazo zorrotzak hemen gaingiroki baizik ez aipatuz, gramatika sortzaile batek *lau* osakin-mota ditu:

$$V = \{a, b, c, \ldots\}$$

a, b, c, eta abar, *irteerazko* elementuak dira: →hitzak, fonemak, edo abar.

$$V = \{A, B, \ldots P\}$$

A, B ... gramatika-→*kategoriak* dira: →izen, →adberbio ... P = «perpausa», hasierako elementua, →abstraktua.

$$R = \{R_1, R_2, R_3 \ldots\ldots R_n\}$$

(→berridazketa-erregelak: berridatz ezazu ezkerraldeko atala, →erregelak eskatzen dituen aldaketen arabera)

1. *Adibidez:*

$G = \langle V_T , V_A , P, R \rangle$

$V_T = \{a, b\}$ (V_T multzoak, kasu erraz honetan, irteerazko elementu pare bat baizik ez du: *a* eta *b)*

$V_A = \{A, B, P\}$ (V_A multzoak, P perpaus abstraktuaz gain, bi gramatika-unitate dauzka: A eta B)

$R = \{R_1, R_2, R_3, R_4, R_5, R_6\}$

Multzo horrek dituen sei erregelak (erregela erdi-thuear Chomskyren gramatiketan) hauek izaki:

R_1 ... P → aA
R_2 ... P → aB
R_3 ... A → aB
R_4 ... A → aA
R_5 ... B → bB
R_6 ... B → b

Lehenengo bost erregelek ez gaituzte sorkuntza-kateatik ateratzen; A eta B elementuak abstraktuak agertzen direlako. *Seigarren* erregelak, berriz, atera egiten gaitu, ateratzen ari garen mailaren arabera, «teorema», «→monema» edo «perpausa» delakoen maila zehatzeraino.

2. *Molda ditzagun* orain gramatika horren araberako *perpaus* batzu. Esate baterako, erabil dezagun R_1 erregela lehendabizi, R_3 segidan, eta R_6 bukatzeko. Zer lortzen dugu?

1) urratsa ematerakoan, hau egingo dugu: ezkerraldean agertzen dena, «p», R_1 erregelaren arabera, «aA» berridatzi:

P → aA

2) urratsa egitean, era berean, ezkerraldean agertzen dena «aA», R_3 erabiliz, «a.aB» berridatziko dugu; hau da, «aaB».

Eta 3) urratsa egitean, bukatzeko, ezkerraldean agertzen dena «aaB», R_6 erabiliz, honela berridatziko dugu: «aa.b»; hau da, *«aab»*.

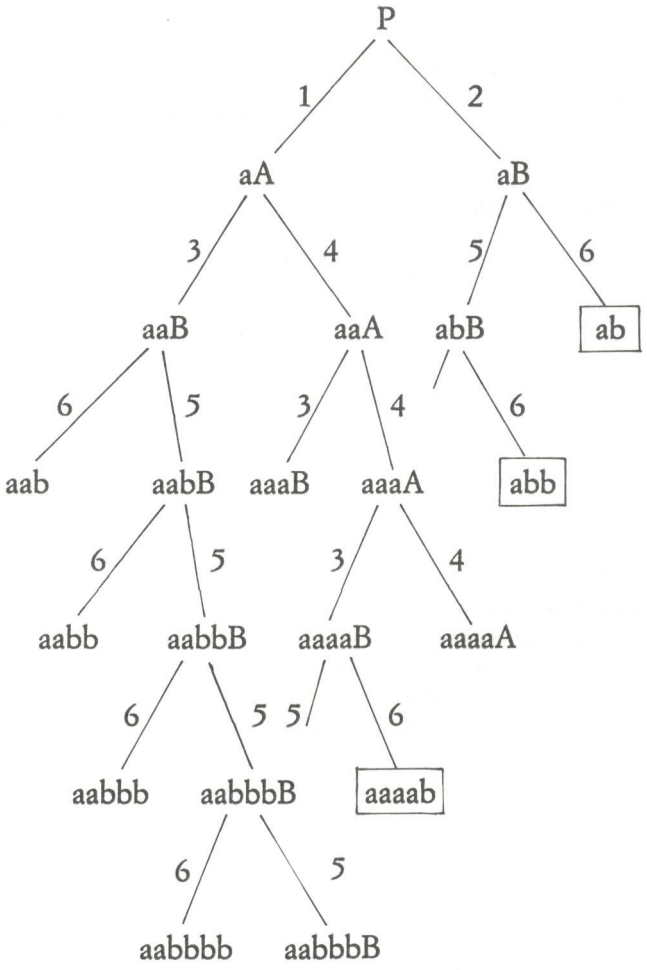
3. Erabil ditzagun orain, bide beretik, eta ilara honetan, erregela hauek: R_2, R_5, R_5 eta R_6. Eta hau dugu:

1 ... R_2 ... P → aB
2 ... R_5 ... aB → abB

3 ... R_5 ... abB → abbB
4 ... R_6 ... abbB → *abbb*

Sei erregela horiek baliatuz, horretara, eta irakurleak egizta dezakeenez, nahi hainbat «perpaus» molda daitezke: *aab, aabb, aabbb, aabbbb, ab, abb, aabb,* eta abar.

Erregela -multzo *finitu* batek, beraz, perpaus- multzo *infinitu* bat eman dezake. Eta hiztunaren *sormena* honetara azal liteke Chomsky-ren ustez.



Hona hemen arestian emandako gramatika sortzaileari dagokion ZUHAITZ SINTAGMATIKOA. Aisa ikus daitekeenez, elkarren kide dira biok; baina, askotan, argi-iturri oparoa izan daiteke erregelei dagokien zuhaitz hau marraztea.

BIBLIOGRAFIA

Glotokronologiaz

PENCHOEN, T.: «La Glottochronologie», La Pléiade, La Languge, 865-884 orr.

Zipf-en legeaz

GUIRAUD, P.: La Pléiade, La Languge, 152 eta hurrengoak.
MULLER, Ch.: «Estadística Lingüística», 285 eta hurrengoak.
MARCUS, S.: «Introducción en la Lingüística matemática», 237 eta abar.
GUIRAUD, P.: *Statistique Linguistique,* P.U.F., 1960.

Informazio-Teoriaz

MILLER: *Language and Commanication,* 1951 (frantsesez, P.U.F., 1956).
MARCUS, S.: «Introducción en la Lingüística matemática», 253 eta abar.
SINGH, J.: «Teoría de la Información, del Lenguaje y de la Cibernética», Alianza Universidad, 1972.

Gramatika sortzaileen oinarriez

SERRANO, S.: *Lógica, lingüística y matemáticas,* Anagrama, 1977.
— *Elementos de Lingüística matemática,* Anagrama, 1975.
GLADKIJ, A. V.: *Introducción a la Lingüística matemática,* Planeta, 1972.
GROSS, M.: *Modelos matemáticos en Lingüística,* Gredos, 1976.