15. A nyelv statisztikai tulajdonságai

A nyelvi szerkezet bármelyik szintjén (13. fej.) megszámolhatjuk a különböző egységek előfordulását, és összevethetjük a kapott gyakoriságokat, hogy lássuk, léteznek-e a használatukat szabályozó statisztikai szabályszerűségek. A grammatika, a szókincs, a hangrendszer és az írásrendszer számos aspektusát tanulmányozták már ily módon, s nagyon sok érdekes törvényszerűségre bukkantak. Olyan statisztikai törvényszerűségeket is sikerült találni, amelyek minden nyelvben közösek; ezeket néha statisztikai *törvény*eknek vagy *univerzálé*knak nevezik.

A statisztikai szabályszerűségek függetlenek a beszélőtől vagy az írótól, vagy akár a témától. Míg bizonyos értelemben bármit mondhatunk, amit akarunk, a gyakorlatban nyelvi viselkedésünk szorosan követi a statisztikai elvárásokat. Teljes bizonyossággal állíthatjuk, hogy ha angolul egy *q* betűt írunk, ezt majdnem mindig *u* betű követi (bár vannak kivételek, például az *Iraq* szó és mások). Kevésbé nyilvánvaló ugyan, de ugyanolyan bizonyos, hogy annak, amit mondunk, egy kicsit több mint 60%-a mássalhangzókból áll, és némileg kevesebb mint 40%-a magánhangzókból. Minden köznapi beszédben használt szótagszerkezetnek körülbelül az egyharmada mássalhangzó–magánhangzó–mássalhangzó szekvenciából áll, mint az angol *cat* (macska) szóban. A nyelv 50 leggyakrabban használt szava fogja kitenni körülbelül 45%-át mindannak, amit leírunk.

Az a figyelemre méltó ezekben az adatokban, hogy kommunikáció közben nem ügyelünk ezen statisztikai tulajdonságok érvényesítésére. Ezt nem lehetne megtenni. Ugyanakkor e mögöttes szabályszerűségek mindenféle tudatos erőfeszítés nélkül előállnak beszédünk vagy írásunk bármely nagyobb mintájában. Ezeket a szabályszerűségeket és korlátozó tényezőiket a nyelvstatisztika vizsgálja.

Betűgyakoriság

A nyelvbeli statisztikai szabályszerűségek egyik legegyszerűbb megnyilvánulása az ábécé betűinek előfordulási gyakorisága. A következőkben bemutatunk egy gyakorisági listát, amelyhez az amerikai angol különböző stílusainak összehasonlító vizsgálata alapján jutottak (forrás: A. Zettersten 1969, 21. p.): (a) újságriport, (b) vallásos írás, (c) tudományos írás, (d) regény. A rangsorátlagok, melyek 15 különböző kategóriájú szöveg leírásán alapulnak, összesen több mint egymillió szót számításba véve, az (e) oszlopban találhatók. Az (f) oszlop a Samuel Morse (1791–1872) által a morzeábécé létrehozásakor használt rangsort adja meg. Az ő gyakorisági sorrendje a nyomda által használt különböző betűtestek mennyiségén alapult, lásd a (g) oszlopot. A (h)

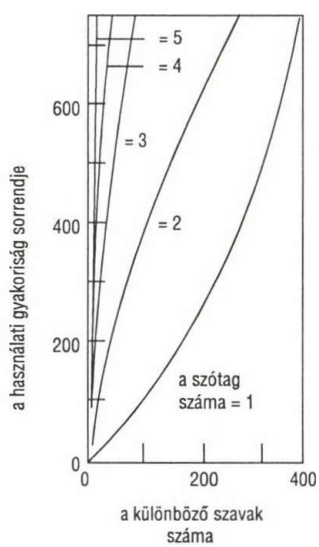
20-as szó-„slágerlista"

A táblázat a 20 leggyakrabban előforduló szót mutatja be egy francia, német és angol újságokon végzett vizsgálat alapján. (Forrás: P. M. Alexejew et al., 1968) Az összehasonlítás kedvéért megadjuk a London–Lund beszélt nyelvi korpuszban szereplő leggyakoribb szavakat (510. p.). A beszéd és írás közötti különbségtétel jelentősége nyilvánvaló: figyeljük meg az *I, yes* és *well* szavak gyakoriságát a beszélt angolban, és a DDR szó (Deutsche Demokratische Republik = Német Demokratikus Köztársaság) előfordulását a német listán.

Rangsor	Francia	Német	Írott angol	Beszélt angol
1	de	der	the	the
2	le (ne.)	die	of	and
3	la (ne.)	und	to	I
4	et	in	in	to
5	les	des	and	of
6	des	den	a	a
7	est	zu	for	you
8	un (ne.)	das	was	that
9	une (ne.)	von	is	in
10	du	für	that	it
11	que (nm.)	auf	on	is
12	dans	mit	at	yes
13	il	sich	he	was
14	à	daß	with	this
15	en	dem	by	but
16	ne	sie	be	on
17	on	ist	it	well
18	qui	im	an	he
19	au	eine	as	have
20	se	DDR	his	for

ne. = névelő; nm. = névmás

Egyszótagú vagy többszótagú?



A leggyakoribb szavak egyszótagúak. Ez világosan kiderül egy telefonbeszélgetéseket vizsgáló tanulmányból. Nagyon kevés 3 vagy ennél több szótagú volt a 800 leggyakrabban előforduló szó között.

(Forrás: N. R. French et al., 1930)

oszlopban a magyar nyelv hasonló adatait adjuk meg. Füredi–Kelemen, szerk. (1989) alapján, mely a hatvanas és hetvenes évek szépirodalmának félmilliónál több szövegszavát dolgozza fel:

(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
e	e	e	e	e	e	12000	e
t	t	t	t	t	t	9000	a
a	i	a	a	a	a	8000	t
o	a	i	o	o	i	8000	l
n	o	o	h	i	n	8000	n
i	n	n	n	n	o	8000	s
s	s	s	i	s	s	8000	k
r	r	r	s	r	h	6400	o
h	h	h	r	h	r	6200	m
l	l	l	d	l	d	4400	r
d	d	c	l	d	l	4000	i
c	c	d	u	c	u	3400	g
m	u	u	w	u	c	3000	á
u	m	m	m	m	m	3000	é
f	f	f	c	f	f	2500	d
p	p	p	g	p	w	2000	b
g	y	g	f	g	y	2000	v
w	w	y	y	w	g	1700	h
y	g	b	p	y	p	1700	j
b	b	w	b	b	b	1600	ö
v	v	v	k	v	v	1200	f
k	k	k	v	k	k	800	p
j	x	x	j	x	q	500	u
x	j	q	x	j	j	400	ő
q	q	j	z	q	x	400	ó
z	z	z	q	z	z	200	c
							ü
							í
							ú
							ű
							w

ZIPF TÖRVÉNYEI

A nyelv jelentős statisztikai törvényszerűségeit kimutató egyik első kutató George Kingsley Zipf amerikai filológus (1902–1950) volt. Legismertebb „törvénye" szerint egy szó gyakorisági listán elfoglalt helye és a szövegbeli használatának gyakorisága közötti kapcsolatot egy állandó jellemzi. Ha ellenőrizni akarjuk a törvény érvényességét, a következő műveleteket kell elvégeznünk:

1. Számoljuk meg a különböző szavak összes előfordulását egy szövegben – például, a *the* (ne.) szó 364 eset, az *is* (létige) szó 251, a *table* (asztal) szó 4 és így tovább.
2. Tegyük őket csökkenő gyakorisági sorrendbe, és adjunk mindegyiknek sorszámot – (1) *the* 364, (2) *is* 251, (3) *of* 166 stb.
3. Szorozzuk meg a sorszámot (r) a gyakorisággal (f), és az eredmény megközelítőleg állandó lesz (C).

Például, az alábbi lista a London–Lund beszélt nyelvi korpusz (510. p.) egyik kategóriájának a 35., 45., 55., 65. és 75. leggyakrabban előforduló szavát adja meg. Az értékek mindegyike 30 000 körül van.

r	×	f	=	C
35	very (nagyon)	836	=	29 260
45	see (lát)	674	=	30 330
55	which (amelyik, amely)	563	=	30 965
65	get (vesz, kap stb.)	469	=	30 485
75	out (ki, kint)	422	=	31 650

Más szóval a kapcsolat fordítottan arányos, és ezt mindig érvényesnek gondolták, függetlenül a témától, szerzőtől vagy bármi más nyelvészeti változótól. Később azonban kimutatták, hogy a kapcsolat nem áll fenn a leggyakoribb és a legkevésbé gyakori szavak esetében. Ugyanabban a korpusz-

Szótárak

Fogjunk egy szótárt, és számoljuk meg az egyes szavak alcímszók szerinti jelentéseit! Az olyan szavak száma (n), amelyeknek bizonyos számú jelentésük van (m), fordítottan arányos a jelentések számának négyzetével ($n \times m^2 = C$).

Nézzünk meg egy szöveget, bármilyen szöveget

Nézzünk meg egy tetszőleges nyelvű szöveget, és számoljuk meg benne a szavakat! Rangsoroljuk őket csökkenő gyakoriság szerint. Statisztikai előrejelzés azt mutatja, hogy az első 15 szó adja a szöveg 25%-át. Az első 100 szó már a szöveg 60%-át teszi ki, az első 1000 a 85%-át, az első 4000 pedig a 97,5%-át. Ezek az arányok kisebb mintákban jelentősen eltérő képet mutatnak.

Hosszúság/gyakoriság összefüggések

A szótagszám és az előfordulási gyakoriság kapcsolatát mutatja az alábbi táblázat, közel 11 millió német szó vizsgálata alapján.

(Forrás: F. W. Kaeding, 1898)

A szó szótagjainak száma	A szavak előfordulásának száma	Az összes előfordulás százalékában
1	5 426 326	49,76
2	3 156 448	28,94
3	1 410 494	12,93
4	646 971	5,93
5	187 738	1,72
6	54 436	0,50
7	16 993	
8	5 038	
9	1 225	
10	461	
11	59	0,22
12	35	
13	8	
14	2	
15	1	

ban például a leggyakoribb szó, *I* (én) 5920 alkalommal fordul elő (r × f = 5920), és a 100. szó, *he's* (ő+létige, hn.) 363-szor (r × f = 36 300). A minta mérete ugyancsak kritikus tényező.

Mindazonáltal, a szógyakoriság „szabályos görbéje", képletbe foglalva f × r = C, érdekes megfigyelést jelent a nyelvi mintázatokról. Sőt hasonló jellegű görbét találtak más nyelvekben is. Például egy francia szógyakoriságot bemutató könyv szerint a 100-ik szót 314 alkalommal használták (= 31 400), a 200-ikat 158-szor (= 31 600), az 1000-iket pedig 31-szer (= 31 000). A magyarban a 100-ik szó 516-szor (= 51 600), a 200-ik 277 alkalommal (= 55 400), az 1000-ik pedig 57-szer (= 57 000) fordul elő.

Egyéb összefüggések

Zipf azt is kimutatta, hogy fordított összefüggés van a szó hossza és gyakorisága között. Az angolban például a gyakran használt szavak többsége egy szótagú. Ugyanez az összefüggés érvényesül még olyan jellegzetesen „több szótagos" szókészletű nyelvekben is, mint a német vagy a magyar. Ez a hatás magyarázza, miért rövidítünk le szavakat, ha gyakoriságuk növekszik, például az angol *microphone* (mikrofon) szó rutinszerű lerövidítése *mike*-ra a rádiósok beszédében, vagy a magyarban: *magnetofon – magnó*. Emellett hatékony kommunikációs elvnek is tűnik, hogy a sokat használt szavak rövidek, a ritka szavak pedig hosszúak legyenek.

Zipfet, aki a „legkisebb erőfeszítés" elve alapján kívánta megmagyarázni a hangok és szavak sokfélesége és egyformasága között fennálló nyilvánvaló egyensúlyt, nagyon érdekelték az olyan tényezők, mint a kommunikáció hatékonysága és egyszerűsége. Minél egyszerűbb egy hang, és rövidebb egy szó, annál gyakrabban használjuk őket. Ez a magyarázat számos problémát rejt magában (például: miként mérhető, számszerűsíthető az egyes hangok képzésével járó „erőfeszítés", valamint a szabály fentebb említett kivételei), és napjainkra egy, a valószínűségszámítás elméletére épülő, hagyományosabb felfogás lett elfogadott.

George Kingsley Zipf (1902–1950)

Szótagok

Vegyük valamilyen beszélt angol nyelvi anyag magnófelvételét, és írjuk át fonetikai átírással! Jelöljük meg a szótaghatárokat! Azt láthatjuk, hogy 12 szótag teszi ki a beszédanyag 25%-át: /ðə/, /əv/, /ɪn/, /ænd/, /ɪ/, /ə/, /tʊ/, /ɪŋ/, /ə-/, /rɪ/, /ɪt/, /ðæt/ (lásd a II. Függeléket a fonetikai átírás szimbólumaihoz). A szöveg fele csak 70 különböző szótagot használ. Ha viszont a 90%-át tekintjük a szövegnek, már több mint 1300 szótagtípusra van szükségünk. A *the* (a, az) szótag egyedül az összes beszélt nyelvi szótag 7%-át adja; átlagosan 14 szótagonként fordul elő. (Forrás: G. Dewey, 1923)