A Zipf- és a Heaps-törvény

A korpuszokban előforduló különböző szavak számának és gyakoriságának leírására szolgál a tapasztalatokon alapuló Zipf- és Heaps-törvény (Heaps, 1978).

Rendezzük sorba a korpuszunk szavait csökkenő gyakoriság szerint. Elöl tehát a leggyakoribb, hátul a legritkábban előforduló — tipikusan egyszer, nyelvészeti műszóval *hapax*nak nevezett — szavak lesznek. A Zipf-törvény azt mondja ki, hogy a lista i. elemének gyakorisága:

$$R_i \sim \frac{1}{i^\alpha},$$

ahol α egy a korpuszra jellemző, 1 körüli konstans.

Más szavakkal ez azt jelenti, hogy ha log–log skálán ábrázoljuk i-t és R_i-t, akkor a pontok jó közelítéssel egy egyenesen helyezkednek el. A törvény a leggyakrabban és a legritkábban előforduló szavakra kevésbé pontos.

A törvény érdekes következménye, hogy ha egy korpuszból csak a leggyakoribb szavakat tartjuk meg, a többit töröljük, a korpusz nagy része akkor is megmarad. Példa: ha 30 000 különböző szavunk van, és $\alpha = 1{,}1$, és a 15 000 leggyakrabban előforduló szót tartjuk meg, akkor a korpuszban levő szavak több, mint 96%-át megtartottuk, miközben a szótárat a felére csökkentettük. Ha csak az 1 000 leggyakoribb szót tartjuk meg, akkor ugyanez az arány majdnem 80%, miközben a szótárat a 30-adára csökkentettük!

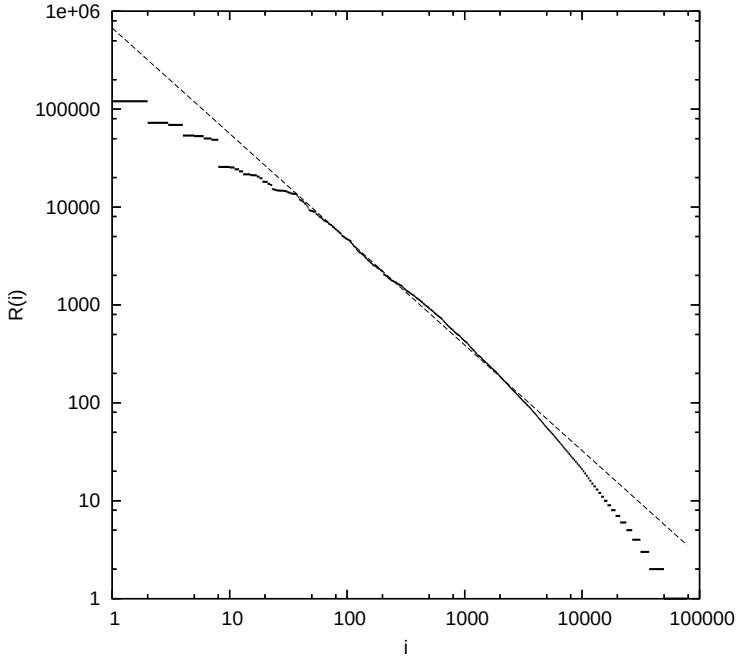
A Heaps-törvény szerint egy N szóból álló korpusz esetén a különböző szavak száma:

$$V \approx K \cdot N^\beta$$

ahol:

- K: a szövegtől függő konstans, tipikusan: $10 \leq K \leq 100$
- β: a szövegtől függő konstans, angol nyelvre $0{,}4 \leq \beta \leq 0{,}6$, magyar nyelvre inkább 0,6 és 0,7 között van. Kisebb N-ek esetén nagyobb β szükséges, kb. $N > 150\,000$-től stabilizálódik.

A törvény következménye, hogy egy korpuszban a különböző szavak száma új dokumentumok hozzáadásával folyamatosan növekszik, például az elírásoknak, a tulajdonneveknek és az új szavaknak köszönhetően. Szerencsére ez a növekedés csak szublineáris, és az új szavak nyilván nem lesznek gyakori szavak. A növekedés következménye, hogy minél több dokumentumunk van, annál ritkább lesz a szó-dokumentum mátrix.



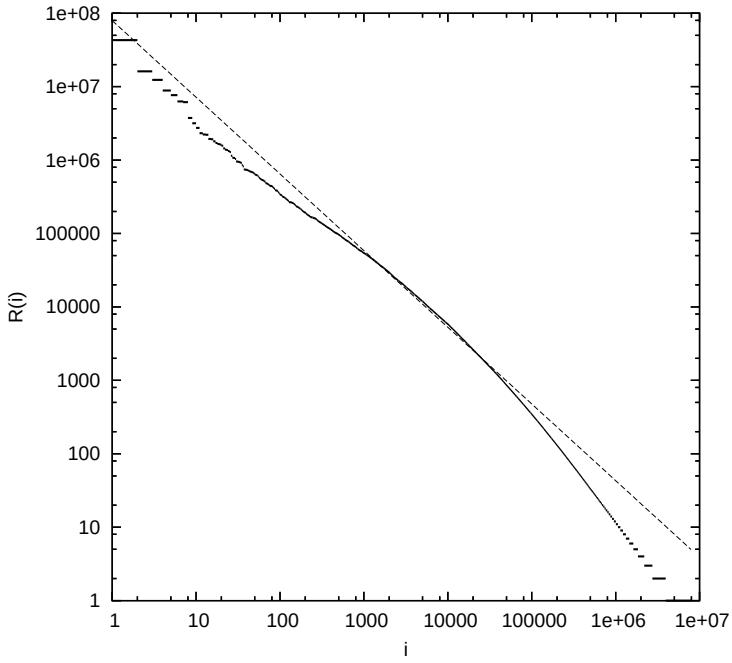
1. ábra. Zipf-törvény a gyakorlatban a Reuters-21578 angol nyelvű korpuszon

Példa: $N = 3 \cdot 10^6$ és $\beta = 0{,}7$ esetén $K = 10$ esetén $V = 3{,}4 \cdot 10^5$, azaz ha az elejétől olvassuk a korpuszt, úgy tűnik, mintha minden kilencedik szó új lenne.

Az 1. és 2. ábrákon a Zipf-törvény megvalósulása látható a gyakorlatban. Az 1. ábra a teljes Reuters-21578 korpusz, míg a 2. ábra Web2.2 korpusz (Halácsy et al, 2004) 4%-os szűrésével kapott dokumentumgyűjtemény alapján készült. A Web2.2 korpusz a jelenlegi legnagyobb, $1{,}48 \cdot 10^9$ szót tartalmazó magyar nyelvű korpusz, amely a magyar web alapján készült 2003-ban. A 4%-os szűrés azt jelenti, hogy a weboldalak közül csak azokat tartották meg, amelyek esetén az alkalmazott helyesírás-ellenőrző a szavaknak legalább a 96%-át helyesnek ítélte.

Forrás:

H. S. Heaps. *Information Retrieval — Computational and Theoretical Aspects.* Academic Press, 1978.



2. ábra. Zipf-törvény a gyakorlatban a magyar nyelvű Web2.2-es korpuszon

P. Halácsy, A. Kornai, L. Németh László, A. Rung, I. Szakadát, V. Trón. Creating open language resources for Hungarian. In *Proc. of LREC'04, 4th Int. Conf. on Language Resources and Evaluation*, 2004.