Przetwarzanie języków naturalnych - analiza statystyczna języka węgierskiego

Jakub Grzyb, Paweł Froń

8 października 2024

1 Wstęp

Na podstawie próbki tekstów w języku węgierskim, zawierających w sumie około 46 tyś. słów, przeprowadzono analizę statystyczną tego języka. W szczególności, zweryfikowano prawo Zipfa, wyznaczono listy 10%, 20%, 30%, 40% i 50% najczęściej występujących słów, oraz sporządzono graf przedstawiający powiązania między wyrazami.

2 Prawo Zipfa

Tabela 1 zawiera wybarne wyrazy wraz z ich rangą i częstotliwością. Można zaobserwować, że iloczyn rangi i częstotliwości jest z w przybliżeniu taki sam, co jest zgodne z prawem Zipfa.

3 Listy wyrazów

Lista

3.1 Najczęstsze 10%

a, és, az

3.2 Najczęstsze 20%

a, és, az, nem, hogy, is, egy, de, csak, van, mint, már

3.3 Najczęstsze 30%

a, és, az, nem, hogy, is, egy, de, csak, van, mint, már, ez, meg, most, ha, vagy, úgy, még, volt, kell, mikor, ki, vagyok, s, el, igen, se, ahogy, azt, én, ami, egész, valami, hát, volna, fel, magam, aztán, itt, azt, mi, olyan, mintha, nagyon, talán, aki

Tabela 1: Wybrane wyrazy wraz z ich rangą i częstotliwością

Słowo	Ranga (r)	Częstotliwość (f)	r * f
a	1	2963	2963
és	2	1067	2134
az	3	1037	3111
nem	4	941	3764
hogy	5	854	4270
is	6	469	2814
egy	7	418	2926
de	8	346	2768
csak	9	298	2682
van	10	286	2860
volt	20	160	3200
azt	30	113	3390
itt	40	98	3920
tudom	50	85	4250
akkor	60	73	4380
felé	70	62	4340
rá	80	53	4240
ahol	90	46	4140
más	100	40	4000

3.4 Najczęstsze 40%

a, és, az, nem, hogy, is, egy, de, csak, van, mint, már, ez, meg, most, ha, vagy, úgy, még, volt, kell, mikor, ki, vagyok, s, el, igen, se, ahogy, azt, én, ami, egész, valami, hát, volna, fel, magam, aztán, itt, ezt, mi, olyan, mintha, nagyon, talán, aki, amit, mégis, tudom, lehet, nincs, mert, ott, nagy, kis, két, szó, hiszen, akkor, ő, minden, jó, lesz, jól, éppen, e, után, át, felé, majd, be, megint, ember, mindig, így, valamit, pedig, milyen, rá, semmi, mit, előtt, látszik, ne, le, kicsit, nekem, egyszer, ahol, reggel, azért, tanár, neki, soha, végre, egyik, alatt, ilyen, más, látom, fél, érzem, fejem, úr, sok, furcsa, eszembe, múlva, rám, elég, első, láttam, szép, nélkül, tudja, miért, bennem, különben, maga, kedves, sem, na, másik, voltam, közben, beteg, három, vele, dolog, velem, róla, először, tovább, egyre, hosszú, olivecrona, gyorsan, arra

3.5 Najczęstrze 50%

a, és, az, nem, hogy, is, egy, de, csak, van, mint, már, ez, meg, most, ha, vagy, úgy, még, volt, kell, mikor, ki, vagyok, s, el, igen, se, ahogy, azt, én, ami, egész, valami, hát, volna, fel, magam, aztán, itt, ezt, mi, olyan, mintha, nagyon, talán, aki, amit, mégis, tudom, lehet, nincs, mert, ott, nagy, kis, két, szó, hiszen, akkor, ő, minden, jó, lesz, jól,

éppen, e, után, át, felé, majd, be, megint, ember, mindig, így, valamit, pedig, milyen, rá, semmi, mit, előtt, látszik, ne, le, kicsit, nekem, egyszer, ahol, reggel, azért, tanár, neki, soha, végre, egyik, alatt, ilyen, más, látom, fél, érzem, fejem, úr, sok, furcsa, eszembe, múlva, rám, elég, első, láttam, szép, nélkül, tudja, miért, bennem, különben, maga, kedves, sem, na, másik, voltam, közben, beteg, három, vele, dolog, velem, róla, először, tovább, egyre, hosszú, olivecrona, gyorsan, arra, észre, hozzá, néhány, szinte, régi, ennek, ezek, óra, új, vagyunk, benne, látni, magát, nap, kellene, őket, bár, persze, valaki, mondja, holnap, fehér, jön, kérem, vannak, utolsó, rólam, néha, fáj, azzal, tudok, egyetlen, pontosan, tessék, ezzel, rögtön, baj, jut, sincs, orvos, egyszerre, móni, tudtam, rajta, óta, ma, érdekes, áll, közt, feleségem, túl, való, akinek, akartam, akar, délután, össze, fog, senki, mindjárt, este, vissza, előbb, erre, tud, ám, inkább, jutott, hallom, akit, nemcsak, igaz, no, csakugyan, hol, semmit, megy, értem, történt, kellett, ezért, ebben, együtt, lehetett, fiatal, erről, emlékszem, következő, lett, fogok, megyek, később, lám, hallottam, mindent, komoly, jobban, több, míg, isten, hét, tehát, lenni, orvosi, annak, tíz, magamban, valahol, beszél, te, jobb, rossz, pár, tart, szóval, gyula, idő, mely, ismeretlen, tudni, lassan, kezd, szabad, ismerem, szót, fölött, hat, jött, daganat, fekete, történik, aminek, veszem, ezen, ismerős, külön, utána, hideg, igazán, hanem, finom, ezúttal, többé, óvatosan, négy, óriási, őnagysága, mindenki, körül, szívesen, emberi, sokszor, élet, ekkor, csodálkozva, dolgot, lehetne, mondta, nini, műtét, biztosan, mondani, amiket, oldalt, ugyan, idegen, rövid, pillanat, valóság, azon, annyi, sokat, cini, tetszik, természetesen, órakor, rajtam, halkan, fontos, amiben, nyilván, lefelé, sokáig, bécsi, köszönöm, oda, ajtó, szemben, elé, bizony, teljesen, szegény, ezelőtt, perc, mögött, külső, kerül, barátom, pillanatban, újra, éreztem, délelőtt, egyedül, azonban, engem, tesz, hirtelen, kezdtem, pillanatra, hozzám, tudod, figyel, apró, vizsgálat, ugye, mellett, ide, igyekszem, szóba, bizonyos, nálam, amire, gondoltam, eddig, szerény, amivel, nehéz, utcán, szalad, vonat, otthon, előre, szemem, abba, végig, életemben, másnap, tettem, határozottan, miről, vége, mondom, látok, egyelőre, tartott, él, üres, ó, kék, magyar, fekszem, került, érdekli, akár, esetleg, egészen, akik, halk, hittem

4 Graf

Graf zawiera informację o tym, jakie wyrazy ze sobą sąsiadują oraz ile razy. W sporządzonym grafie zdecydowanie wyróżniają się następujące wyrazy:

a - 2071 połączeń,
és - 776,
az - 573,
nem - 519,

hogy - 491,
is - 351,
egy - 295

Można z tego wywnioskować, że są to słowa o istotnej funkcji gramatycznej.