# COMPARAREA PROBABILITĂŢILOR ELEMENTELOR TEXTULUI

***V.Bobicev***
*Universitatea Tehnica a Moldovei*

## INTRODUCERE

Procesarea automată a limbajului natural este o ramificare a lingvisticii computaţionale. Trebuie de subliniat, totuşi, că în timp ce lingvistica computaţională abordează mai mult caracteristicile lingvistice ale limbajului, procesarea limbajului natural (PLN) se ocupă în special de aspectele informatice ale limbajului. Pe parcursul dezvoltării PLN s-a recurs la aplicarea a numeroase metode ale teoriei informaţiei, teoriei probabilităţilor şi statisticii. Se consideră în acest moment că folosirea metodelor statisticii matematice în prelucrarea limbajului natural reprezintă un domeniu de vârf în PLN.

Noţiunea de bază pentru metodele statistice este probabilitatea elementelor textului în limbaj natural. Majoritatea metodelor statistice sunt bazate pe o presupunerea că apariţia oricărui element în text (litera, silabă, cuvânt) este aleatoare, supusă legilor statisticii, şi, deci, este posibil de estimat probabilitatea apariţiei elementului dat în text cu o acurateţe stabilită.

Aplicarea metodelor statistice se bazează pe utilizarea corpusurilor (singular: corpus, plural: corpora), care sunt colecţii on-line de texte. În studiul de faţă sunt folosite patru corpusuri de texte române. Descrierea corpusurilor este dată în tabelul 1.

**Tabelul 1.** Descrierea corpusurilor, folosite în studiul dat.

| **denumirea scurtă a corpus-ului** | **Descrierea** | **sursa** | **volum (în Kb)** | **numărul total de caractere** |
|---|---|---|---|---|
| **Hot** | hotărîrile Curţii de Apel şi Recurs a Moldovei | http://moldova.wjin.net | 4 553 | 4 511 057 |
| **RL** | arhiva site-ului *România Literară* | www.romlit.ro | 49 712 | 48 648 093 |
| **Ad** | ziarul on-line *Adevărul* | http://www.adevarulonline.ro | 15 624 | 14 787 010 |
| **EZ** | ziarul on-line *Evenimentul Zilei* | http://www.expres.ro | 57 076 | 50 995 252 |

În [4] este prezentat un studiu profund al comportamentului matematic al limbii române scrise prin metode statistice. Au fost analizate structurile de litere, digrame, trigrame şi tetragrame de litere. În acest scop a fost creat un corpus lingvistic în baza căruia a fost construit un ansamblu statistic care conţine: (a) estimarea probabilităţii unui eveniment; (b) testul de apartenenţa a probabilităţii la un interval; (c) testul de egalitate între două probabilităţi. Ca rezultat, au fost obţinute probabilităţi reprezentative a literelor şi secvenţelor de litere. A fost justificată staţionaritatea limbii române la nivel de litere. Însă comparaţia între textele literare şi ştiinţifice arată diferenţe între probabilităţile literelor.

Scopul studiului de faţă este verificarea egalităţii probabilităţilor literelor şi secvenţelor de litere în corpus-uri diferite. În scopul acesta au fost efectuate următoarele cercetări: (1) au fost estimate probabilităţile literelor şi secvenţelor de litere în corpus-uri date; (2) au fost înaintate ipoteze de egalitate a probabilităţilor estimate; (3) ipotezele au fost verificate folosind testul cu ***z***-criteriu.

## 1. ESTIMAREA PROBABILITĂŢILOR

Modelarea statistică a textului poate fi efectuată prin mai multe căi. Spre exemplu, o variabilă aleatoare poate fi considerată distanţa între apariţii consecutive a unui cuvânt în text. Însă, apariţiile cuvintelor consecutive pot fi considerate aleatoare, dar nu pot fi considerate independente

unul de altul. Deci, în scopul obţinerii valorilor independente în [4] a fost propusă următoare metodă. Textul este eşantionat periodic cu o perioadă suficient de mare astfel încât putem considera că observaţiile succesive sunt independente statistic. În [4] a fost propus un pas de eşantionare de 200 caractere. Prima mulţime este formată din cuvintele cu numărul de ordine 1, 201, 401, 601, … în text. A doua mulţime este formată din cuvintele cu numărul de ordine 2, 202, 402, 602, … şi aşa mai departe. Astfel sunt formate 200 de mulţimi cu lungimea N/200, unde N este numărul total de litere în corpus.

În aşa mod au fost formate mulţimi în care elementele sunt absolut independente. Însă noi nu putem tot atât de categoric să afirmăm existenţa dependenţei între mulţimile date.

O problemă importantă în calculele statistice este stabilirea volumului necesar de texte pentru obţinerea probabilităţilor statistic reprezentative. Există mai multe abordări ale problemei date. Uneori cerinţele de reprezentativitate staistică sunt înlocuite cu criteriul posibilităţii de realizare. În astfel de cazuri mărimea mulţimii este determinată de factori externi şi anume de volumul textelor disponibile. Altă problemă care apare în calculul probabilităţilor este estimarea raportului optimal între volumul mulţimilor şi numărul lor. În (4), cum s-a spus, au fost formate 200 mulţimi reieşind din condiţia de Moivre-Laplace:

$$N \cdot p' \cdot (1-p') >> 10 \qquad (1)$$

unde

N – volumul mulţimii;

P’ - frecvenţa relativă.

În [4] a fost înaintată condiţia:

$$N \cdot p' \cdot (1-p') \geq 20 \qquad (2)$$

Rescriem formula (2):

$$N = 20 / (p' \cdot (1-p')) \qquad (3)$$

Calculăm o valoare a frecvenţei minimale care satisface condiţia de Moivre-Laplace, reieşind din volumul mulţimii N = 225 553. Rezultatele pentru patru corpusuri folosite sunt prezentate în Tabelul 2.

În baza calculelor putem face concluzia că rezultatele sunt reprezentative pentru toate litere în afară de «q», «w», «y», «k», frecvenţa cărora este prea mică. Pentru secvenţe din două şi trei litere respectiv aproape 32% de secvenţe din două litere şi

**Tabelul 2.** Frecvenţa minimală care satisface condiţia de Moivre-Laplace.

| corpus | număr total de litere | frecvenţa minimală |
|---|---|---|
| **Hot** | 4 511 057 | 0,0009 |
| **RL** | 48 648 093 | 0,00008 |
| **Ad** | 14 787 010 | 0,0003 |
| **EZ** | 50 995 252 | 0,00008 |

12% de secvenţe din trei litere au frecvenţa mai mare decît cea minimală calculată. Deci, o parte din secvenţele de litere pot fi folosite în studiul următor.

Pentru estimarea probabilităţilor literelor corpusurile au fost transformate în forma următoare. Toate textele corpusului se unesc într-un text comun de mărimea corpusului dat pentru eşantionarea prin corpusul întreg şi formarea mulţimilor în baza corpusului. Din texte se elimină toate caracterele nealfabetice, deci, în texte rămân numai literele, blancurile şi cratimele, în total 33 de caractere diferite. Toate literele se transformă în minuscule. Textul corpusului obţinut după transformarea descrisă este eşantionat şi sunt estimate probabilităţile literelor ca media aritmetică a valorilor obţinute pentru fiecare mulţime:

$$P' = \left( \sum_{i=1}^{N} p_i' \right) / N \qquad (4)$$

unde

N – numărul total de mulţimi;

$p_i'$ - valoarea frecvenţei relative obţinute pentru fiecare mulţime;

$$p_i' = m_i / M \qquad (5)$$

$m_i$ – numărul de succese (apariţii a elementului dat) pentru fiecare mulţime;

M – volumul mulţimii.

## 2. ÎNAINTAREA ŞI ESTIMAREA IPOTEZELOR DE EGALITATE A PROBABILITĂŢILOR

Probabilităţile literelor în text sunt valori aleatoare şi nu pot fi comparate direct. În astfel de cazuri există metoda numită ‘verificarea ipotezei despre variabila aleatoare’. În studiul de faţă folosim un tip de ipoteze, şi anume, ipoteza de verificare a egalităţii între două probabilităţi pentru două variabile aleatoare.

Dacă dispunem de două mulţimi de date experimentale în baza cărora am aflat estimaţiile probabilităţilor $p'_1$ şi $p'_2$. Urmărim să stabilim dacă

două estimaţii $p'_1$ şi $p'_2$ provin din aceeaşi probabilitate teoretică, respectiv dacă $p_1 = p_2$ .

Procedura de test este următoarea:

- Se formulează cele două ipoteze statistice, ipoteza nulă H0 şi ipoteza alternativă H1:

H0 : cele două probabilităţi teoretice sunt egale, $p_1 = p_2$ ;

H1 : cele două probabilităţi teoretice sunt diferite, $p_1 \neq p_2$ ;

- Se alege gradul de semnificaţie statistică $\alpha$.

- Se construieşte o valoare de test ***z*** conform [1]:

$$z = \frac{p'_1 - p'_2}{\sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}} \qquad (6)$$

Întrucît $p'_1$ şi $p'_2$ sunt necunoscute se consideră

$$p_1 = p_2 \approx (m_1 + m_2) / (N_1 + N_2) \qquad (7)$$

În aceste condiţii valoarea de test ***z*** devine:

$$z = \sqrt{\frac{N_1 + N_2}{N_1 N_2}} \frac{m_1 N_2 - m_2 N_1}{\sqrt{(m_1 + m_2)(N_1 + N_2 - m_1 - m_2)}} \qquad (8)$$

Ipoteza nulă H0 va fi acceptată (se va considera că probabilităţile sunt egale, $p_1 = p_2$) dacă şi numai dacă $|z| \leq z_{\alpha/2}$ ($z_{\alpha/2}$ corespunde pragului de semnificaţie statistică $\alpha$ ales; toate rezultatele experimentale sunt obţinute cu 1- $\alpha$ = 0,95 atunci $z_{\alpha/2} = 1{,}96$)

În caz contrar se respinge ipoteza de egalitate a celor două probabilităţi pentru pragul de semnificaţie statistică $\alpha$ ales.

Aceasta procedură de test este folosită pentru compararea între ele a patru corpusuri de texte.

## 3. VERIFICAREA IPOTEZELOR ÎNAINTATE

Având probabilităţile literelor estimate pentru fiecare corpus înaintăm ipoteza că probabilităţile acestora sunt egale.

Deci, înaintăm două ipoteze:

H0 : probabilităţile literelor în corpusuri diferite sunt egale ;

H1 : probabilităţile literelor în corpusuri diferite sunt diferite ;

Alegem gradul de semnificaţie $\alpha$ = 0,5, atunci 1- $\alpha$ = 0,95, respectiv $z_{\alpha/2} = 1{,}96$.

Apoi calculăm valoarea criteriului ***z*** pentru perechi de corpusuri. Compărând valoarea obţinută cu valoarea corespunzătoare gradului de semnificaţie ales acceptăm sau respingem ipoteza nulă.

În procesul comparaţiei a apărut o dificultate. Nu am putut compara toate corpusuri între ele din cauză că corpusurile date conţin un număr diferit de litere. Corpusul **RL** conţine toate 31 litere române, corpusul **Hot** practic nu conţine litera “â”, corpusuri **EZ** şi **Ad** sunt scrise numai cu litere engleze, toate semnele diacritice sunt omise. Din cauza aceasta corpusurile au fost grupate în perechi **RL** cu **Hot** şi **EZ** cu **Ad**. Perechile acestea s-au comparat, rezultatele sunt prezentate în tabelul 3.

**Tabelul 3.** Valorile ***z***-criteriului pentru literele corpusurilor.

| **Hot şi RL** | | **EZ şi Ad** | |
|---|---|---|---|
| **litere** | **valorile *z*-criteriului** | **litere** | **valorile *z*-criteriului** |
| e | 28,05 | a | 24,06 |
| i | 12,63 | i | 37,11 |
| a | 38,99 | e | 21,39 |
| r | 49,36 | t | 35,61 |
| n | 17,38 | r | 39,89 |
| t | 18,90 | n | 18,33 |
| u | 4,16 | u | 9,96 |
| c | 31,03 | l | 10,64 |
| l | 19,24 | c | 2,44 |
| o | 11,26 | s | 7,73 |
| s | 12,69 | o | 0,30 |
| d | 32,29 | d | 15,79 |
| ă | 0,08 | p | 14,51 |
| m | 28,34 | m | 8,99 |
| p | 71,28 | f | 9,73 |
| î | 9,57 | v | 0,27 |
| ş | 39,56 | b | 14,23 |
| v | 2,95 | g | 6,29 |
| f | 1,15 | z | 5,09 |
| ţ | 40,42 | - | 6,58 |
| b | 31,41 | h | 46,04 |
| g | 35,49 | j | 16,01 |
| z | 20,60 | x | 4,94 |
| - | 66,04 | | |
| h | 1,90 | | |
| â | 40,47 | | |
| x | 80,06 | | |
| j | 79,72 | | |

Valorile ***z***-criteriului prezentate în tabel sunt cu mult mai mari decât valoarea care corespunde pragului de semnificaţie, prin urmare noi respingem ipoteza nulă că probabilităţile literelor estimate în baza corpusurilor de texte diferite sunt egale şi acceptăm ipoteza alternativă că probabilităţile literelor estimate în baza corpusurilor de texte diferite sunt diferite. Necatând la unele excepţii (spre exemplu, pentru litera "h": $z = 1{,}90 \Rightarrow p^h_{Hot} = p^h_{RL}$) putem constata că probabilitatea literelor în texte diferite este diferită.

În [3] a fost dat un exemplu de comparare a corpusurilor în baza cuvintelor. În [2] au fost comparate două corpusuri de texte engleze: american şi britanic. În baza calculelor autorii au ajuns la concluzie că versiunea americană a limbii engleze se deosebeşte de cea britanică. În [3] s-a efectuat un alt experiment. Autorul a împărţit un corpus în două părţi egale şi le-a comparat. El a arătat că diferenţa între părţile unui corpus a fost aproape la fel de mare ca şi diferenţa între corpusurile britanic şi cel american. Prin urmare el a făcut concluzia că engleza americană şi cea britanică se deosebesc nu mai mult decât două genuri diferite a limbii engleze care în general este foarte variată în diferite domenii: limba literară, limbajul presei, limbajul ştiinţific ş.a. Un corpus general al limbii este ca regulă format din documente din diferite domenii. Deci, el este foarte eterogen şi practic nu este posibil de vorbit de o probabilitate unică pentru corpusul în întregime.

Prin urmare noi am comparat părţile corpusurilor având ca scop să aflăm cît de omogene sunt corpusurile noastre şi cît de real este de vorbit de o probabilitate unică pentru un corpus. Posibil că şi în cazul nostru corpusurile sunt prea eterogene şi nu le putem compara direct. Pentru aceasta am împărţit fiecare corpus în două părţi egale şi am comparat părţile una faţă de cealaltă. Experimentele au fost efectuate asupra corpusurilor **EZ** şi **Ad**. Rezultatele calculelor sunt expuse în tabelul 4.

**Tabelul 4.** Valorile ***z***-criteriului pentru literele corpusurilor.

| **Părţile corpusului Ad** | | | **Părţile corpusului EZ** | |
|---|---|---|---|---|
| **litere** | **valorile *z*-criteriului** | | **litere** | **valorile *z*-criteriului** |
| a | 4,06 | | a | 1,64 |
| i | 1,56 | | i | 0,77 |
| e | 4,98 | | e | 0,20 |
| t | 6,30 | | t | 0,85 |
| r | 7,21 | | r | 1,48 |
| n | 1,15 | | n | 0,22 |
| u | 2,37 | | u | 1,04 |
| c | 5,27 | | l | 1,05 |
| s | 4,13 | | c | 2,39 |
| l | 7,11 | | s | 0,90 |
| o | 3,24 | | o | 1,63 |
| d | 0,18 | | d | 0,85 |
| p | 3,40 | | p | 0,39 |
| m | 0,06 | | m | 0,06 |

După cum se observă din tabel, valorile ***z***-criteriului, fără îndoială, sunt mai joase decât la compararea diferitor corpusuri. Pentru corpusul **EZ** noi putem să constatăm cu siguranţă că probabilităţile literelor în majoritatea cazurilor sunt egale în ambele părţi ale corpusului. Pentru **Ad** nu putem spune acest lucru atât de categoric, fiindcă pentru un număr mare de litere valorile criteriului sunt mai mari decât valoarea care corespunde pragului de semnificaţie ce nu ne permite să acceptăm ipoteza nulă. Totuşi, valorile criteriului sunt cu mult mai mici decât în tabelul precedent, deci putem spune că probabilităţile sunt cu mult mai apropiate.

Toate textele în corpusuri sunt scrise de autori diferiţi. Posibil că diferenţa între probabilităţile literelor este cauzată şi de diferenţa între autorii textelor în corpus.

La fel au fost comparate probabilităţile perechilor şi secvenţelor din trei litere în text. Fragmentele rezultatelor calculelor pentru perechile literelor a corpusurilor **Ad** şi **EZ** şi a părţilor corpusului **EZ** sunt expuse în tabelul 5.

**Tabelul 5.** Valorile ***z***-criteriului pentru perechile literelor corpusurilor.

| **Ad şi EZ** | | | **Părţile corpusului EZ** |
|---|---|---|---|
| **perechi de litere** | **valorile *z*-criteriului** | | **valorile *z*-criteriului** |
| a_ | 25,89 | | 2,42 |
| e_ | 24,21 | | 0,26 |
| i_ | 5,09 | | 0,93 |
| in | 52,76 | | 0,20 |
| _d | 3,68 | | 0,54 |
| _a | 7,60 | | 2,32 |
| _c | 6,18 | | 1,20 |
| _s | 6,89 | | 1,28 |
| re | 33,28 | | 1,59 |
| a_ | 22,34 | | 0,01 |
| … | … | | … |
| oi | 1,47 | | 1,80 |
| gr | 1,10 | | 0,76 |
| ze | 0,27 | | 0,79 |

| | | | |
|---|---|---|---|
| ng | 0,04 | | 0,49 |
| rs | 0,06 | | 0,32 |
| ex | 0,60 | | 0,47 |
| ju | 0,41 | | 2,33 |
| rc | 1,55 | | 0,93 |
| cl | 3,87 | | 0,27 |
| ab | 0,32 | | 1,08 |

Pentru secvenţe de litere tendinţa rămâne aceiaşi – pentru două corpusuri diferite ipoteza nulă nu poate fi acceptată, pentru părţile unui corpus putem spune că probabilităţile sunt egale.

Trebuie de menţionat că dacă pentru litere nu era o problemă compararea fiecărei litere, pentru secvenţe de litere situaţia este cu mult mai complicată. Numărul secvenţelor de litere este destul de mare chiar dacă considerăm numai secvenţele cu frecvenţa înaltă. Din cauza aceasta în [4] au fost propuse două tipuri de comparaţii între secvenţele literelor:

- comparaţii între secvenţele literelor ca atare;
- comparaţii între secvenţele literelor cu acelaşi rang.

Pentru compararea în baza rangului valorile ***z***-criteriului în general sunt mai mici decât pentru compararea în baza secvenţelor literelor ca atare. În afară de aceasta, secvenţele cu frecvenţa medie în text formează grupe care au frecvenţă egală. Anume secvenţele acestea formează un grup comparativ mare de elemente care cel mai bine este descris cu formula lui Zipf [4] şi cel mai bine se supun formalizării în modelele statistice.

**Tabelul 6.** Valorile z-criteriului pentru secvenţe din trei litere din texte.

| **Ad şi EZ** | | | **Părţile corpusului EZ** |
|---|---|---|---|
| **secvenţe din trei litere** | **valorile z-criteriului** | | **valorile z-criteriului** |
| _de | 3,47 | | 0,33 |
| de_ | 18,06 | | 0,12 |
| _in | 6,94 | | 2,14 |
| in_ | 4,06 | | 1,69 |
| _ca | 20,07 | | 1,44 |
| ul_ | 2,66 | | 0,83 |
| re_ | 3,88 | | 0,07 |
| … | … | | … |
| ala | 1,25 | | 0,14 |
| e_n | 1,57 | | 0,06 |
| scu | 1,93 | | 0,21 |
| ci_ | 2,01 | | 0,27 |
| oar | 1,78 | | 0,41 |
| _ar | 1,18 | | 0,04 |
| _ro | 0,95 | | 0,01 |

După cum se observă din tabelul 6, valorile criteriului sunt încă mai joase decât pentru secvenţe din două litere. Presupunem, că micşorarea diferenţei între corpusuri în conformitate cu creşterea lungimii secvenţelor este datorită faptului că corespunzător creşterii lungimii secvenţelor creşte numărul lor şi, respectiv, creşte numărul secvenţelor cu frecvenţa medie. Prin urmare, numărul secvenţelor cu frecvenţa înaltă scade brusc. Anume secvenţele acestea au nişte frecvenţe individuale care se încadrează foarte slab în legea lui Zipf şi se modelează mai greu în modelele statistice. Însă, pe de altă parte, anume secvenţele acestea sunt reprezentative statistic şi sunt importante pentru studiul nostru. De aceea noi facem concluzia că frecvenţele secvenţelor de litere în corpusuri diferite diferă considerabil şi modelul creat în baza textelor asemănătoare va lucra cu mult mai bine decât modelul creat în baza textelor din alt domeniu.

Altă serie de experimente a fost realizată de noi pe baza cuvintelor. Ca şi pentru litere am comparat corpusurile între ele şi părţile unui corpus. Reieşind din faptul că numărul de cuvinte este foarte mare, am executat comparaţia cuvintelor cu acelaşi rang. În tabelul 7 sunt prezentate unele fragmente de rezultate obţinute.

**Tabelul 7.** Valorile ***z***-criteriului pentru cuvintele textelor.

| **Ad şi EZ** | | **Părţile corpusului EZ** | | **Hot şi RL** | |
|---|---|---|---|---|---|
| **cuvinte** | **valorile z-criteriului** | **cuvinte** | **valorile z-criteriului** | **cuvinte** | **valorile z-criteriului** |
| de | 7,737 | de | 0,661 | de | 0,157 |
| in | 10,543 | in | 0,639 | şi | 21,809 |
| a | 21,052 | a | 1,509 | în | 14,029 |
| si | 11,114 | si | 1,084 | a | 11,500 |
| la | 7,857 | la | 1,432 | la | 5,956 |
| … | … | … | … | … | … |
| rang 101 | 1,226 | rang 101 | 0,347 | rang 101 | 0,142 |
| 102 | 0,397 | 102 | 0,394 | 102 | 0,046 |
| 103 | 0,808 | 103 | 0,064 | 103 | 0,043 |
| 104 | 0,855 | 104 | 0,054 | 104 | 0,315 |
| 105 | 0,852 | 105 | 0,330 | 105 | 0,233 |

Datele din tabel ne arată că rezultatele obţinute pentru cuvinte sunt încă mai evidente. Pentru părţile unui corpus valorile ***z***-criteriului sunt mai mari decât valoarea care corespunde pragului de semnificaţie numai pentru şase cuvinte. Pentru restul cuvintelor putem accepta ipoteza de egalitate a probabilităţilor. Pentru două corpusuri diferite, însă, diferenţa între probabilităţi este foarte mare.

În tabel este reflectat faptul că 10-15 din cele mai frecvente cuvinte se află aproape pe aceleaşi poziţii în dicţionarele de frecvenţe pentru toate corpusurile. Cuvintele mai puţin frecvente deja diferă mult în corpusuri diferite. Comparaţia a fost executată pentru 150-350 cuvinte mai frecvente în dependenţă de corpusuri din care cele mai mari diferenţe arătau primele 100 de cuvinte.

Următorul experiment executat este comparaţia probabilităţilor perechilor de cuvinte. Unele fragmente din rezultatele obţinute sunt prezentate în tabelul 8.

**Tabelul 8.** Valorile ***z***-criteriului pentru perechile cuvintelor din texte.

| **Ad şi EZ** | | | **Părţile corpusului EZ** |
|---|---|---|---|
| **perechi de cuvinte** | **valorile *z*-criteriului** | | **valorile *z*-criteriului** |
| de la | 1,353 | | 0,477 |
| a fost | 16,391 | | 2,980 |
| au fost | 1,522 | | 0,591 |
| pe care | 1,846 | | 1,083 |
| de ani | 1,070 | | 0,220 |
| … | … | | … |
| rang 101 | 0,356 | | 0,053 |
| 102 | 0,318 | | 0,413 |
| 103 | 0,737 | | 0,242 |
| 104 | 0,558 | | 0,185 |
| 105 | 1,132 | | 0,112 |

După cum se vede din tabel, practic nu este înregistrată diferenţa între probabilităţile perechilor de cuvinte cum pentru părţile unui corpus aşa şi pentru corpusuri diferite. Posibil, rezultatul acesta este datorit faptului că practic nu există perechi cu frecvenţa înaltă, care provoacă diferenţa la nivelul cuvintelor. Unica excepţie este perechea de cuvinte: 'a fost' pentru care rezultatele se deosebesc de celelalte.

## CONCLUZII

Modelele statistice ale textului sunt pe larg folosite în procesarea limbajului natural. Noţiunea de bază pentru modelele statistice este probabilitatea elementelor textului. Probabilităţile se estimează folosind un corpus de texte. Caracteristicile corpusului au o influenţă imensă asupra caracteristicilor modelului creat. De aceea este foarte important de analizat caracteristicile statistice a corpusurilor şi de comparat cu caracteristicile textelor la care va fi aplicat modelul creat. În studiul dat am comparat probabilităţile literelor şi cuvintelor în corpusuri diferite. Scopul comparării este determinarea egalităţii probabilităţilor, şi justificarea folosirii modelului statistic creat în baza unui corpus pentru textele din alt domeniu. Au fost comparate litere, secvenţe din două şi trei litere, cuvinte, secvenţe din două cuvinte. Pentru compararea probabilităţilor a fost folosită metoda ipotezelor statistice cu ***z***-criteriu.

Rezultatele experimentelor au arătat că în toate cazurile ipoteza de egalitate a probabilităţilor elementelor textului pentru corpusuri diferite nu poate fi acceptată şi, prin urmare este respinsă. Aceasta înseamnă că probabilităţile elementelor textului nu sunt egale în corpusuri diferite. Deci modelul statistic, creat în baza unui corpus practic reflectă caracteristicile statistice textelor din corpusul dat şi poate fi efectiv aplicată la texte de acelaşi tip. Aplicarea modelului la texte de alt tip nu va fi efectivă.

În general putem face concluzia că fiecare tip de texte are nevoie de un model statistic particular atât la nivel de cuvinte cât şi la nivel de litere.

## *Bibliografia*

***1. Devore, J.*** *Probability and Statistics for Engineering and the Sciences, second edition, Books/Cole Publishing Company, Monterey, California, 1987.*

***2. Hofland, K., Johansson, S.*** *Editors. Word Frequenceis in British and American English. The Norwegian Computing Centre for the Humanities, Bergen, Norway. 1982.*

***3. Kilgarriff, A.*** *Using word frequency lists to measure corpus homogeneity and similarity between corpora. In Proceedings of ACL-SIGDAT Workshop on very large corpora, Hong Kong, 1997.*

***4. Vlad, A., Mitrea, A., i Mitrea, M.*** *Limba română scrisă ca sursă de informaţie. Paideia, România. 2005.*

***Recomandat spre publicare: 11.05.2006***