

ORTOGONALITATEA ALFABETELOR

Ion Ivan

ionivan@ionivan.ro

Daniel Milodin

daniel.milodin@ie.ase.ro
Academia de Studii Economice, București

Marius Popa

marius.popa@ie.ase.ro

Rezumat. Se definește conceptul de ortogonalitate a textelor. Sunt evidențiate structurile sintactice care formează entitățile text. Pentru alfabetele considerate în analiza ortogonalității, se construiește un sistem de referință. Măsura în care simbolurile alfabetelor cifrelor romane, slav și grec sunt diferite este evidențiată prin intermediul metricilor de ortogonalitate. De asemenea, se prezintă și se aplică un algoritm de cuantificare a ortogonalității. Implementarea algoritmului are loc într-un produs software, rezultatele obținute din rularea sa fiind prezentate în cadrul lucrării.

Cuvinte cheie: ortogonalitate, alfabet, metrici, sistem de referință.

1. Ortogonalitatea – caracteristică de calitate a textelor

Fie a_1, a_2, \dots, a_n simbolurile elementare ale alfabetului A .

Cu simbolurile alfabetului A se construiesc cuvintele c_1, c_2, \dots, c_p ale vocabularului V .

Datele reprezintă șiruri de cuvinte puse în corespondență cu obiecte, acțiuni, fenomene, reprezentări ale lumii reale. Datele se construiesc după reguli sau se generează.

Elementele din lumea reală diferă unele de altele. Similar, simbolurile alfabetelor și cuvintele construite diferă unele de altele. În lumea reală, se identifică domenii de activitate. Un domeniu este fie o arie geografică, un set de activități omogene, o colecție de obiecte, o colectivitate formată din persoane care au elemente comune, fie un sistem funcțional.

Totalitatea cuvintelor cu ajutorul cărora se descrie complet, corect, sugestiv un domeniu se numește *tezaur*.

Un *text* este o înșiruire de cuvinte. Lungimea textului este dată de numărul de cuvinte care intră în alcătuirea sa. Cuvintele sunt separate printr-un simbol din alfabet ales drept delimitator, și care nu mai intră în alcătuirea cuvintelor din vocabular.

Se consideră mulțimea textelor $\tau = \{T_1, T_2, \dots, T_{nt}\}$, unde T_i este un text format din cuvintele $C_1^i, C_2^i, \dots, C_l^i$ din vocabularul V . Se notează f_h^i frecvența de apariție a cuvântului c_h în textul T_i .

De exemplu, se consideră alfabetul $A = \{a, b, c, d, x\}$, separatorul fiind caracterul x . Pe baza alfabetului A , se construiește vocabularul $V = \{aa, abc, bac, dda, ccb, adb, abdc\}$. Se consideră următoarele texte:

| |
|--|
| $T_1 = \langle aaxabcxabdcxaaxaaxabcxaa \rangle$ |
| $T_2 = \langle bacxddaxbacxbacxdda \rangle$ |
| $T_3 = \langle abdcxabdcxabdcxaaxaaxaaxaa \rangle$ |

Lungimile textelor ca număr de cuvinte sunt: $L(T_1) = 7$ cuvinte, $L(T_2) = 5$ cuvinte și $L(T_3) = 8$ cuvinte. Frecvențele de apariție ale cuvintelor din vocabularul V sunt evidențiate în tabelul 1.

Tabelul nr. 1 Frecvențele de apariție ale cuvintelor din textele T_1, T_2 și T_3

| Cuvinte \ Texte | T_1 | T_2 | T_3 |
|-----------------|-------|-------|-------|
| aa | 4 | 0 | 4 |
| abc | 2 | 0 | 0 |
| bac | 0 | 3 | 0 |
| dda | 0 | 2 | 0 |
| ccb | 0 | 0 | 0 |
| adb | 0 | 0 | 0 |
| abdc | 1 | 0 | 3 |

Textul T_i este format din mulțimea cuvintelor diferite $V_D^i = \{C_1^i, C_2^i, \dots, C_m^i\}$, care au frecvențele de apariție $f_1^i, f_2^i, \dots, f_m^i$. Astfel: $V_D^1 = \{aa, abc, abdc\}$, $V_D^2 = \{bac, dda\}$ și $V_D^3 = \{aa, abdc\}$.

Dintre caracteristicile de calitate, ortogonalitatea este importantă pentru că stă la baza originalității textelor. Ortogonalitatea se evidențiază la nivel de alfabet, cuvinte, vocabular prin intermediul frecvențelor de apariție.

Se consideră textele T_i și T_j . Ortogonalitatea celor două texte este evidențiată pe baza următoarelor elemente:

- lungimea textelor exprimată ca număr de cuvinte;
- vocabularele textelor, V_i și V_j ;
- frecvențele de apariție a cuvintelor în vocabulare;
- intersecții de vocabulare.

Dacă $V_i \cap V_j = \emptyset$ textele sunt originale din punctul de vedere al cuvintelor, dar nu și din punctul de vedere al semnificației conținutului.

În analiza caracteristicii de ortogonalitate, se urmărește maniera în care textele T_i și T_j sunt ortogonale. Ortogonalitatea vizează măsura în care textele diferă ca formă de prezentare și conținut.

Două texte sunt ortogonale dacă ele nu au nimic în comun. Elementele e_i și e_j din T_i , respectiv T_j sunt ortogonale când cuvintele situate pe pozițiile similare sunt distincte. Identitatea a două texte se definește în mod contrar ortogonalității. Astfel, textele sunt identice dacă au același volum, iar valorile de pe pozițiile similare sunt egale. Identitatea se definește sub următoarea expresie analitică:

$$I = I - O$$

unde:

- I – identitatea textelor;
- O – ortogonalitatea textelor.

Textele T_i și T_j sunt identice dacă:

- vocabularele sunt identice: $V_i \subset V_j$ și $V_j \subset V_i$;
- frecvențele de apariție ale cuvintelor sunt identice: $f_{ik} = f_{jk}$;
- pozițiile cuvintelor sunt identice în textele T_i și T_j .

Se identifică două forme ale ortogonalității: internă și externă.

Ortogonalitatea internă vizează legăturile care există între construcțiile sintactice ale unei entități text.

Ortogonalitatea externă are în vedere diferențele care există între texte. Lucrările elaborate într-un anumit domeniu sunt 100% ortogonale dacă textele au în comun numai cuvintele de specialitate, iar celelalte cuvinte sunt foarte diferite ca frecvențe de apariție și poziție.

Este mai important ca lucrările să fie ortogonale, și nu total diferite. Dacă lucrările sunt total diferite, înseamnă că nu aparțin aceluiași domeniu.

Evidențierea ortogonalității dintre două texte ale unei colectivități are loc prin intermediul indicatorului I_i . Acesta are valorile cuprinse în intervalul $[0, 1]$. Valorile extreme au următoarele semnificații:

- $I_i = 0$ – elementele sunt ortogonale, adică nu au nimic în comun;
- $I_i = 1$ – elementele sunt identice, adică nu diferă prin nici o caracteristică.

Caracteristica de ortogonalitate este strâns legată de conceptul de distanță. O distanță mai mare înseamnă un nivel mai ridicat de ortogonalitate.

Ortogonalitatea textelor joacă un rol important în dezvoltarea unor lucrări cu un nivel ridicat al calității. Identificarea unică a unui text dintr-o colectivitate are loc pe baza conceptului de amprentă.

Amprenta se construiește prin efectuarea unui sondaj statistic. Un eșantion reprezintă o parte a unei colectivități care este analizată. Construirea și determinarea mărimii unui eșantion se realizează prin intermediul metodelor și tehnicilor specifice sondajului statistic. Pe baza analizei unui eșantion, se desprind concluzii, idei care sunt valide și la nivel de colectivitate. Acest lucru se realizează numai dacă se asigură reprezentativitatea eșantionului.

2. Sistem de referință

Entitatea este o noțiune primară. O entitate este descrisă printr-o serie de caracteristici. Se consideră caracteristicile C_1, C_2, \dots, C_m cu care se descriu cel puțin două entități. Ortogonalitatea măsoară cât de diferite sunt entitățile. Ortogonalitatea este măsura diferenței dintre elementele unei colectivități.

Se consideră un alfabet A format din simbolurile $a_1, a_2, \dots, a_n, A = \{a_1, a_2, \dots, a_n\}$.

În cazul alfabetului, se definește un sistem de referință $R = \{r_1, r_2, \dots, r_k\}$ (figura 1), cu ajutorul căruia se descriu simbolurile:

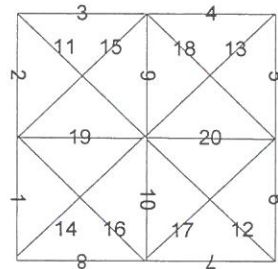


Figura 1. Sistemul de referință

Dacă se consideră alfabetul corespunzător simbolurilor asociate cifrelor romane $B = \{I, V, X, L, C, D, M\}$, folosind sistemul de referință, se obțin reprezentările date în figura 2.

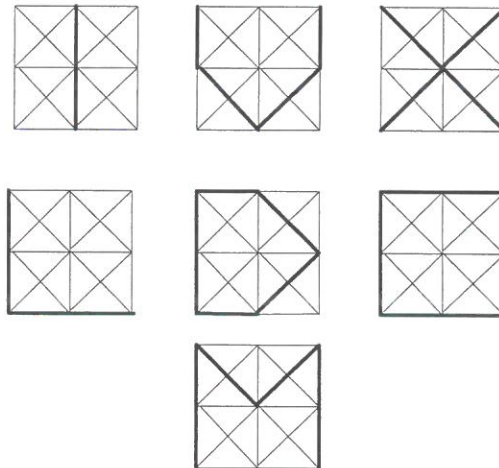


Figura 2. Reprezentarea cifrelor romane cu sistemul de referință

Se observă că simbolurile nu se confundă unele cu altele pentru că sunt diferite. Pentru a cuantifica gradul de diferențiere dintre două simboluri ale unui alfabet, se propune algoritmul:

Pasul 1: se definește un sistem de referință format din segmentele r_1, r_2, \dots, r_k ;

Pasul 2: se asociază câte un rang α_i fiecărui segment r_i din sistemul de referință, obținând perechile (r_i, α_i) ;

Pasul 3: se construiesc simbolurile alfabetului A folosind sistemul de referință;

Pasul 4: se construiește o matrice $M(A_L)$ asociată cifrelor romane astfel încât elementul $x_{ij}=1$ dacă pentru a construi simbolul a_i din alfabetul A_L se utilizează segmentul de referință r_j ;

Pasul 5: se calculează sumele pe coloanele $S_j = \sum_{i=1}^n x_{ij}$, $j = 1, 2, \dots, k$, pentru a totaliza frecvențele de utilizare a segmentelor din sistemul de referință în definirea simbolurilor din alfabet;

Pasul 6: se interschimbează coloanele matricei $M(A_L)$ pentru a obține o dispunere în ordine descrescătoare după totalurile S_j , $j = 1, 2, \dots, k$. Prima coloană corespunde segmentului r_{\max} deoarece totalul său S_{\max} are valoare maximă, iar ultima coloană corespunde segmentului r_{\min} , întrucât totalul S_{\min} este cu valoare minimă;

Pasul 7: se recodifică rangurile segmentelor sistemului de referință, astfel încât noile numere să reflecte frecvențele de utilizare, obținându-se perechile (r_j, β_j) .

Pentru simbolurile alfabetului A_L descrise în figura 2, se construiește matricea frecvențelor sistemului de referință. Aplicarea algoritmului conduce la obținerea noului sistem de referință, tabelul 2.

Tabelul nr. 2 Corespondența poziții vechi – poziții noi

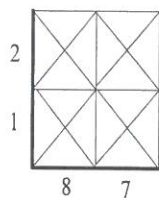
| R | r ₁ | r ₂ | r ₃ | r ₄ | r ₅ | r ₆ | r ₇ | r ₈ | r ₉ | r ₁₀ | r ₁₁ | r ₁₂ | r ₁₃ | r ₁₄ | r ₁₅ | r ₁₆ | r ₁₇ | r ₁₈ | r ₁₉ | r ₂₀ |
|----------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| α | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| β | 2 | 1 | 4 | 10 | 5 | 11 | 6 | 3 | 12 | 13 | 17 | 14 | 8 | 15 | 18 | 16 | 9 | 7 | 19 | 20 |

Pe baza noului sistem de referință, se determină frecvențele din tabelul 3.

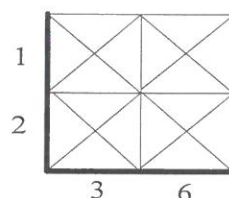
Tabelul nr. 3 Frecvențele noului sistem de referință

| R | r ₁ | r ₂ | r ₃ | r ₄ | r ₅ | r ₆ | r ₇ | r ₈ | r ₉ | r ₁₀ | r ₁₁ | r ₁₂ | r ₁₃ | r ₁₄ | r ₁₅ | r ₁₆ | r ₁₇ | r ₁₈ | r ₁₉ | r ₂₀ |
|---------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| β | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| V | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| X | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| L | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| C | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| M | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

Pentru vechiul sistem de coduri, simbolului L îi corespunde sistemul de referință din figura 3a, și noul sistem de referință din figura 3b.



a) vechiul sistem de ranguri



b) noul sistem de ranguri

Figura 3. Schimbarea rangurilor pentru simbolul L din alfabetul A_L

Pentru a măsura ortogonalitatea dintre simbolul a_i și simbolul a_j trebuie definit un indicator $H(a_i, a_j) \in [0;1]$, astfel încât:

$H(a_i, a_j) = 0$, dacă cele două elemente sunt identice;

$H(a_i, a_j) = 1$, dacă cele două elemente sunt complet diferite.

Pentru caracteristicile de calitate Ch_1, Ch_2, \dots, Ch_m , ortogonalitatea vizează măsura în care elementele a_i și a_j sunt înzestrate cu caracteristici de calitate diferite. Elementele a_i și a_j sunt identice dacă au aceleași niveluri ale caracteristicilor de calitate.

Când elementul a_i este înzestrat cu submulțimea de caracteristici Ch' și elementul a_j este înzestrat cu submulțimea de caracteristici Ch'' și $Ch' \cap Ch'' = \emptyset$ rezultă că elementele a_i și a_j sunt ortogonale dacă sunt privite numai prin prisma acestui mod de a remarca diferențele dintre ele.

Pentru măsurarea ortogonalității dintre simbolurile a_i și a_j ale alfabetului A_L se definește indicatorul cu forma analitică:

$$H(a_i, a_j) = \frac{\sum_{k=1}^m (\bar{\gamma}_{ki} \cap \gamma_{kj}) \cup (\gamma_{ki} \cap \bar{\gamma}_{kj})}{\sum_{k=1}^m (\gamma_{ki} \cup \gamma_{kj})}$$

unde:

- γ_{ki} evedențiază prezența sau absența segmentului r_k în reprezentarea simbolului i din alfabetul A_L ;
- m reprezintă numărul de segmente din sistemul de referință.

Pentru simbolurile care formează setul cifrelor romane folosind sistemul de referință $R = \{r_1, r_2, \dots, r_{20}\}$, se calculează matricea simetrică a ortogonalității dată în tabelul 4.

Tabelul nr. 4 Matricea de ortogonalitate a cifrelor romane

| | C | D | I | L | V | X | M |
|---|------|------|------|------|------|------|------|
| C | 0,00 | 0,50 | 1,00 | 0,33 | 0,89 | 1,00 | 0,80 |
| D | 0,50 | 0,00 | 1,00 | 0,57 | 0,75 | 1,00 | 0,80 |
| I | 1,00 | 1,00 | 0,00 | 1,00 | 1,00 | 1,00 | 1,00 |
| L | 0,33 | 0,57 | 1,00 | 0,00 | 0,86 | 1,00 | 0,75 |
| V | 0,89 | 0,75 | 1,00 | 0,86 | 0,00 | 1,00 | 0,75 |
| X | 1,00 | 1,00 | 1,00 | 1,00 | 1,00 | 0,00 | 0,75 |
| M | 0,80 | 0,80 | 1,00 | 0,75 | 0,75 | 0,75 | 0,00 |

Se observă că cele două alfabete au simboluri care diferă unele de altele. Există simboluri care diferă mai puțin unele de altele, precum C și D, C și L. De asemenea, există simboluri ce diferă foarte mult unele de altele, precum C și V, D și I sau I și C.

Pentru a avea o imagine globală asupra ortogonalității simbolurilor din alfabet, se definește indicatorul de forma:

$$H(A_L) = \frac{n(n-1)}{2} \sqrt{\prod_{i=1}^{n-1} \left(\prod_{j=i+1}^n H(a_i, a_j) \right)}$$

Pentru alfabetul cifrelor romane, $H(A_L) = 0,87$.

Când se construiește un indicator, acesta trebuie să fie: *senzitiv, necatastrofic, necompensatoriu*.

Indicatorul $H(a_i, a_j)$ este *senzitiv* întrucât, pentru variații mici ale structurii simbolurilor, se obțin variații mici ale indicatorului H, iar pentru variații mari ale structurilor se obțin, de asemenea, variații mari ale indicatorului.

Indicatorul $H(a_i, a_j)$ este *necatastrofic* pentru că toate valorile sunt diferite de zero, numitorul nefiind niciodată nul.

Indicatorul este *compensatoriu* pentru multe dintre simbolurile diferite ale alfabetului A_L , obținându-se aceleași valori, tabelul 5.

Tabelul nr. 5 Frecvențele nivelurilor identice pentru matricea $M(A_L)$

| Valoare | Frecvență |
|---------|-----------|
| 1,00 | 20 |
| 0,75 | 8 |
| 0,00 | 7 |
| 0,80 | 4 |
| 0,50 | 2 |
| 0,33 | 2 |
| 0,89 | 2 |
| 0,57 | 2 |
| 0,86 | 2 |

Se pune problema construirii unui indicator de ortogonalitate $H'(a_i, a_j)$, care să amelioreze caracterul compensatoriu reducând numărul valorilor identice ale indicatorului pentru perechi de simboluri diferite.

Fiecărui segment r_1, r_2, \dots, r_k i se asociază perechile p_1, p_2, \dots, p_k , obținând tripletele (r_i, α_i, p_i) , astfel încât $p_i \in [0;1]$ și $\sum_{i=1}^k p_i = 1$.

Se construiește indicatorul:

$$H'(a_i, a_j) = \frac{\sum_{s=1}^k \alpha_{is} \alpha_{js} p_s}{\max\{S'(a_i), S'(a_j)\}}$$

unde:

$$S'(a_i) = \sum_{j=1}^k \alpha_j^2 p_j \delta_j, \quad \delta_j = \begin{cases} 1, & \text{dacă segmentul } r_j \text{ intră în definierea simbolului } a_i \\ 0, & \text{în caz contrar} \end{cases}$$

De exemplu, simbolul L din alfabetul A_L folosind sistemul de referință $R = \{r_1, r_2, \dots, r_{20}\}$, are reprezentarea:

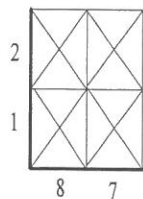


Figura 6. Reprezentarea simbolului L

Tabelul nr. 6 Valorile tabelare ale simbolului L

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|-------|------------|------------|--------|--------------|---------------------------|
| r_1 | 1 | 1 | 0,125 | 1 | 0,125 |
| r_2 | 2 | 1 | 0,1563 | 4 | 1,2504 |
| r_8 | 8 | 1 | 0,0938 | 64 | 48,025 |
| r_7 | 7 | 1 | 0,0625 | 49 | 21,4375 |

Pe baza valorilor din tabelul 6, se obține $H'(L) = 70,83$.

Pentru simbolul I se obțin datele din tabelul 7.

Tabelul nr. 7 Valorile tabelare ale simbolului I

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|----------|------------|------------|--------|--------------|---------------------------|
| r_9 | 9 | 1 | 0,0313 | 81 | 22,81 |
| r_{10} | 10 | 1 | 0,0313 | 100 | 31,3 |

Conform valorilor din tabelul 7, se obține $H'(I) = 54,11$.

Indicatorul de ortogonalitate pentru simbolurile L și I se calculează după relația:

$$H'(I, L) = \frac{\theta}{\max\{H'(I), H'(L)\}}$$

unde θ se obține asemănător cu $H'(I)$, însă doar pentru segmentele comune celor două simboluri. Se obține $H'(I, L) = 0$.

Aplicând această relație pentru toate simbolurile alfabetului se obține matricea $M'(A_L)$. Folosind valorile obținute se completează tabelele valorilor distincte și a frecvențelor de apariție.

Tabelul nr. 8 Matricea $M'(A_L)$

| | C | D | I | L | V | X | M |
|---|--------|--------|--------|--------|--------|--------|--------|
| C | 1,0000 | 0,0945 | 0,0000 | 0,9505 | 0,0028 | 0,0000 | 0,0058 |
| D | 0,0945 | 1,0000 | 0,0000 | 0,0914 | 0,5706 | 0,0000 | 0,0025 |
| I | 0,0000 | 0,0000 | 1,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| L | 0,9505 | 0,0914 | 0,0000 | 1,0000 | 0,0028 | 0,0000 | 0,0058 |
| V | 0,0028 | 0,5706 | 0,0000 | 0,0028 | 1,0000 | 0,0000 | 0,0204 |
| X | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 1,0000 | 0,6121 |
| M | 0,0058 | 0,0025 | 0,0000 | 0,0058 | 0,0204 | 0,6121 | 1,0000 |

Tabelul nr. 9 Frecvențele nivelelor identice ale indicatorului de ortogonalitate

| Valoare | Frecvență |
|---------|-----------|
| 0.0000 | 20 |
| 1.0000 | 7 |
| 0.0028 | 4 |
| 0.0058 | 4 |
| 0.0945 | 2 |
| 0.9505 | 2 |
| 0.0913 | 2 |

| | |
|--------|---|
| 0.5706 | 2 |
| 0.0025 | 2 |
| 0.0204 | 2 |
| 0.6121 | 2 |

Se observă că au crescut valorile indicatorilor și că au scăzut frecvențele lor de apariție. Aceasta înseamnă că, pentru indicatorul H' , caracterul necompensatoriu s-a îmbunătățit.

3. Ortogonalitatea alfabetului slav

Pentru alfabetul slav $A_S = \{\beta, \Gamma, \text{И}, \dots, 3\}$, folosind același sistem de referință se obțin reprezentările date în figura 7.

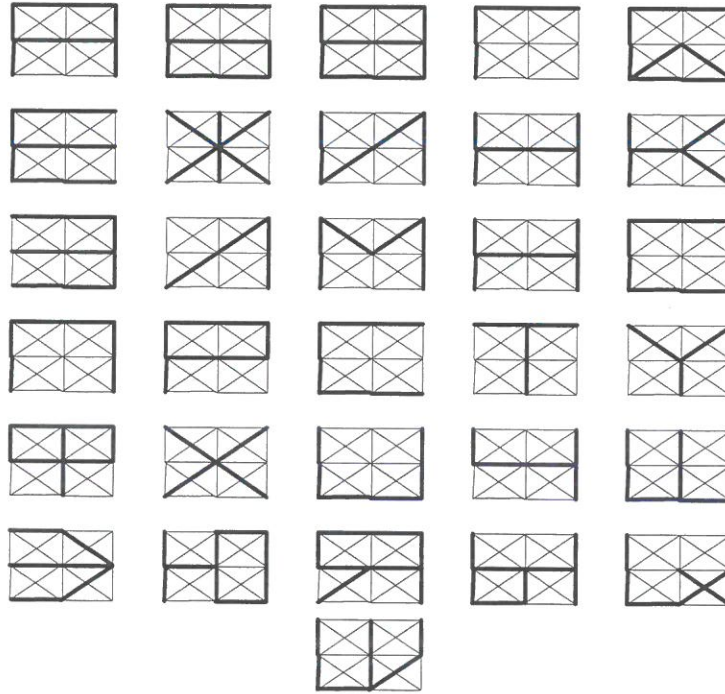


Figura 7. Reprezentarea simbolurilor alfabetului slav

Se observă că simbolurile nu se confundă unele cu altele pentru că sunt diferite. Pentru a cuantifica gradul de diferențiere dintre două simboluri ale alfabetului slav, se aplică algoritmul descris, obținându-se valorile din tabelul 10.

Tabelul 10. Frecvențele noului sistem de referință

| R | r ₁ | r ₂ | r ₃ | r ₄ | r ₅ | r ₆ | r ₇ | r ₈ | r ₉ | r ₁₀ | r ₁₁ | r ₁₂ | r ₁₃ | r ₁₄ | r ₁₅ | r ₁₆ | r ₁₇ | r ₁₈ | r ₁₉ | r ₂₀ |
|---------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| β | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| A | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| V | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| E | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| J | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| I | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Y | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| K | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Z | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| L | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| M | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| N | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| O | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| U | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| F | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |
| I | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| T | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 5 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 6 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |

cu mențiunea că:

Simbol 1: H;

Simbol 2: TS;

Simbol 3: CH;

Simbol 4: SH;

Simbol 5: YU;

Simbol 6: YA;

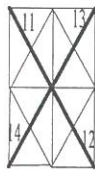
Simbol 7: SCI.

Noul sistem de referință, obținut prin ordonarea frecvențelor, este descris în tabelul 11.

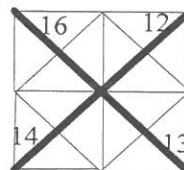
Tabelul 11. Corespondența poziții vechi – poziții noi

| R | r ₁ | r ₂ | r ₃ | r ₄ | r ₅ | r ₆ | r ₇ | r ₈ | r ₉ | r ₁₀ | r ₁₁ | r ₁₂ | r ₁₃ | r ₁₄ | r ₁₅ | r ₁₆ | r ₁₇ | r ₁₈ | r ₁₉ | r ₂₀ |
|---|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| α | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| β | 2 | 1 | 6 | 7 | 3 | 4 | 10 | 9 | 15 | 11 | 16 | 13 | 12 | 14 | 19 | 20 | 17 | 18 | 5 | 8 |

Simbolul H este reprezentat folosind codificarea inițială a sistemului de referință în figura 8a. Folosirea sistemului de referință recodificat în reprezentarea aceluiași caracter este dată în figura 8b.



a) vechiul sistem de ranguri



b) noul sistem de ranguri

Figura 8. Schimbarea rangurilor sistemului de referință pentru simbolul H din alfabetul A_S

Pentru măsurarea ortogonalității dintre simbolurile a_i și a_j ale alfabetului A_S se utilizează indicatorul H(a_i, a_j). Se calculează matricea simetrică de ortogonalitate a simbolurilor alfabetului A_S.

Pentru a avea o imagine globală asupra ortogonalității simbolurilor din alfabet, se propune indicatorul de forma:

$$H(A_S) = \frac{n(n-1)}{2} \sqrt{\prod_{i=1}^{n-1} \left(\prod_{j=i+1}^n H(a_i, a_j) \right)}$$

Pentru alfabetul A_S, al literelor mari din alfabetul rusesc, H(A_S) = 0,66.

Indicatorul este *compensatoriu*, pentru multe dintre simbolurile diferite ale alfabetului A_S obținându-se aceleași valori, tabelul 12.

Tabelul 12. Frecvențele nivelurilor identice pentru matricea M(A_S)

| Valoare | Frecvență | Valoare | Frecvență | Valoare | Frecvență | Valoare | Frecvență |
|---------|-----------|---------|-----------|---------|-----------|---------|-----------|
| 1,00 | 154 | 0,63 | 22 | 0,20 | 8 | 0,13 | 2 |
| 0,50 | 74 | 0,78 | 22 | 0,86 | 8 | 0,22 | 2 |
| 0,80 | 72 | 0,25 | 20 | 0,88 | 8 | 0,10 | 2 |
| 0,60 | 66 | 0,56 | 20 | 0,44 | 8 | 0,85 | 2 |
| 0,75 | 66 | 0,82 | 20 | 0,77 | 6 | 0,11 | 2 |
| 0,67 | 46 | 0,91 | 18 | 0,55 | 6 | 0,58 | 2 |
| 0,83 | 40 | 0,90 | 18 | 0,64 | 6 | 0,79 | 2 |
| 0,00 | 35 | 0,92 | 18 | 0,45 | 4 | 0,36 | 2 |
| 0,40 | 32 | 0,89 | 14 | 0,29 | 4 | 0,43 | 2 |
| 0,70 | 30 | 0,38 | 12 | 0,14 | 4 | | |
| 0,73 | 30 | 0,71 | 12 | 0,17 | 4 | | |
| 0,33 | 22 | 0,30 | 10 | 0,57 | 4 | | |

Se consideră indicatorul de ortogonalitate $H'(a_i, a_j)$ care ameliorează caracterul compensatoriu, reducând numărul valorilor identice ale indicatorului pentru perechi de simboluri diferite. De exemplu, pentru simbolul U din alfabetul A_S folosind sistemul de referință $R = \{r_1, r_2, \dots, r_{20}\}$, prezentat în figura 11.

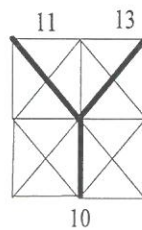


Figura 11. Reprezentarea simbolului U din alfabetul A_S

Tabelul 13. Valorile tabelare ale simbolului U din alfabetul slav

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|----------|------------|------------|--------|--------------|---------------------------|
| r_{10} | 10 | 1 | 0.0345 | 100 | 34,5 |
| r_{11} | 11 | 1 | 0.0197 | 121 | 26,22 |
| r_{13} | 13 | 1 | 0.0296 | 169 | 65,03 |

Pentru valorile date în tabelul 13, se obține $H'(U) = 125,751$.

Pentru simbolul H, se obțin datele din tabelul 14.

Tabelul 14. Valorile tabelare ale simbolului H din alfabetul slav

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|----------|------------|------------|--------|--------------|---------------------------|
| r_{11} | 11 | 1 | 0.0197 | 121 | 26,22 |
| r_{12} | 12 | 1 | 0.0296 | 144 | 51,14 |
| r_{13} | 13 | 1 | 0.0296 | 169 | 65,03 |
| r_{14} | 14 | 1 | 0.0296 | 196 | 81,22 |

Pentru valorile date în tabelul 14, se obține $H'(H) = 223,62$.

Tabelul 15. Valorile tabelare pentru segmentele comune simbolurilor U și H

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|----------|------------|------------|--------|--------------|---------------------------|
| r_{11} | 11 | 1 | 0.0197 | 121 | 26,22 |
| r_{13} | 13 | 1 | 0.0296 | 169 | 65,03 |

Pe baza valorilor tabelare, se calculează $\theta = 91,25$.

Indicatorul de ortogonalitate pentru simbolurile U și H se calculează după relația:

$$H'(U, H) = \frac{\theta}{\max\{H'(U), H'(H)\}}$$

unde θ se obține asemănător cu $H'(U)$, însă doar pentru segmentele comune celor două simboluri. Pe baza valorilor tabelare, $H'(U, H) = 0,40$.

Aplicând această relație pentru toate simbolurile alfabetului se obține matricea $M'(A_S)$. Elementele matricei $M'(A_S)$ sunt valori ale indicatorului H' calculate pentru două caractere ale alfabetului slav. Folosind valorile obținute, se completează tabelul valorilor distincte și frecvențele de apariție ale acestora.

Pe baza matricei de ortogonalitate $M'(A_S)$, se determină frecvențele de apariție pentru valorile indicatorului H' . Se observă că a crescut diversitatea valorilor indicatorului H' și au scăzut nivelurile frecvențelor de apariție. Acest lucru arată că, pentru indicatorul H' , caracterul necompensatoriu s-a îmbunătățit.

4. Ortogonalitatea alfabetului grecesc

Pentru alfabetul grecesc $A_G = \{A, B, \Gamma, \Delta, E, Z, \dots, H\}$, folosind același sistem de referință, se obțin reprezentările date în figura 12.

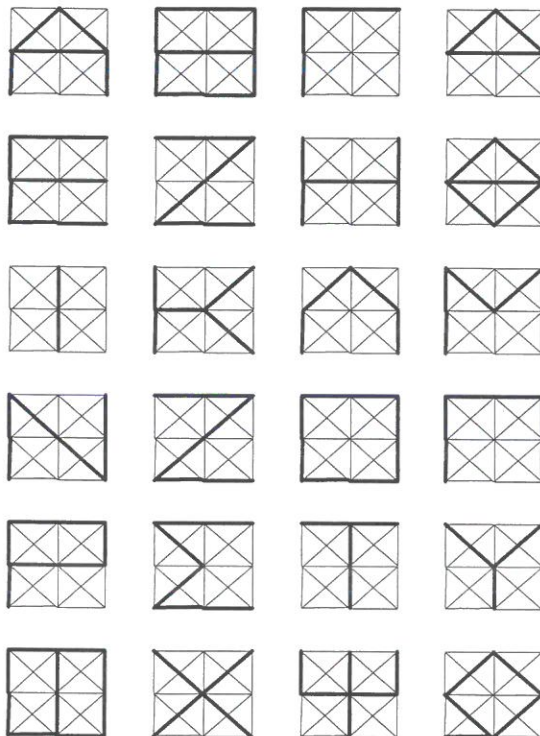


Figura 12. Reprezentarea simbolurilor alfabetului grecesc

Se observă că simbolurile nu se confundă unele cu altele pentru că sunt diferite.

Pentru a cuantifica gradul de diferențiere dintre două simboluri ale unui alfabet, se aplică algoritmul descris, obținându-se frecvențele conform noului sistem de referință.

Tabelul 16. Frecvențele noului sistem de referință

| R | r ₁ | r ₂ | r ₃ | r ₄ | r ₅ | r ₆ | r ₇ | r ₈ | r ₉ | r ₁₀ | r ₁₁ | r ₁₂ | r ₁₃ | r ₁₄ | r ₁₅ | r ₁₆ | r ₁₇ | r ₁₈ | r ₁₉ | r ₂₀ |
|---------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| β | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| A | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| B | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| V | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Z | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| E | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| K | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| M | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| N | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| X | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| O | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| P | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| T | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| U | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| O | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

cu mențiunea că:

Simbol 1: TH;

Simbol 2: PH;

Simbol 3: CH;

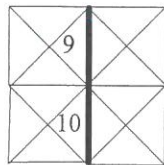
Simbol 4: PS.

Noul sistem de referință obținut prin ordonarea frecvențelor este descris în tabelul 17.

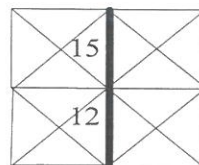
Tabelul 17. Corespondența poziții vechi – poziții noi

| | | | | | | | | | | | | | | | | | | | | |
|---|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| R | r ₁ | r ₂ | r ₃ | r ₄ | r ₅ | r ₆ | r ₇ | r ₈ | r ₉ | r ₁₀ | r ₁₁ | r ₁₂ | r ₁₃ | r ₁₄ | r ₁₅ | r ₁₆ | r ₁₇ | r ₁₈ | r ₁₉ | r ₂₀ |
| α | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| β | 1 | 2 | 3 | 4 | 7 | 5 | 8 | 9 | 15 | 12 | 13 | 18 | 11 | 16 | 14 | 19 | 20 | 17 | 6 | 10 |

Simbolul I este reprezentat folosind codificarea inițială a sistemului de referință în figura 13a. Folosirea sistemului de referință recodificat în reprezentarea aceluiași caracter este dată în figura 13b.



a) vechiul sistem de ranguri



b) noul sistem de ranguri

Figura 13. Schimbarea rangurilor sistemului de referință pentru simbolul I din alfabetul A_G

Pentru măsurarea ortogonalității dintre simbolurile a_i și a_j ale alfabetului A_G se utilizează indicatorul H(a_i, a_j). Se calculează matricea simetrică de ortogonalitate a simbolurilor alfabetului A_G, luate două câte două.

Pentru a avea o imagine globală asupra ortogonalității simbolurilor din alfabet, se definește indicatorul:

$$H(A_G) = \frac{n(n-1)}{2} \sqrt{\prod_{i=1}^{n-1} \left(\prod_{j=i+1}^n H(a_i, a_j) \right)}$$

Pentru alfabetul A_G, al literelor mari din alfabetul grecesc, H(A_G) = 0,75.

Indicatorul este *compensatoriu*, pentru multe dintre simbolurile diferite ale alfabetului A_G obținându-se aceleași valori, tabelul 18.

Tabelul 18. Frecvențele nivelurilor identice pentru matricea M(A_G)

| Valoare | Frecvență | Valoare | Frecvență | Valoare | Frecvență | Valoare | Frecvență |
|---------|-----------|---------|-----------|---------|-----------|---------|-----------|
| 1,00 | 200 | 0,86 | 14 | 0,90 | 8 | 0,38 | 4 |
| 0,75 | 56 | 0,88 | 14 | 0,73 | 6 | 0,10 | 2 |
| 0,80 | 42 | 0,78 | 10 | 0,85 | 6 | 0,11 | 2 |

| | | | | | | | |
|------|----|------|----|------|---|------|---|
| 0,50 | 36 | 0,40 | 10 | 0,20 | 6 | 0,42 | 2 |
| 0,67 | 34 | 0,71 | 10 | 0,29 | 6 | 0,43 | 2 |
| 0,60 | 30 | 0,91 | 10 | 0,63 | 6 | 0,25 | 2 |
| 0,00 | 27 | 0,64 | 8 | 0,89 | 4 | 0,58 | 2 |
| 0,83 | 24 | 0,33 | 8 | 0,56 | 4 | 0,92 | 2 |
| 0,82 | 16 | 0,70 | 8 | 0,30 | 4 | | |

Se consideră indicatorul de ortogonalitate $H'(a_i, a_j)$ care ameliorează caracterul compensatoriu, reducând numărul valorilor identice ale indicatorului pentru perechi de simboluri diferite.

De exemplu, pentru simbolul U din alfabetul A_G folosind sistemul de referință $R = \{r_1, r_2, \dots, r_{20}\}$, prezentat în figura 16.

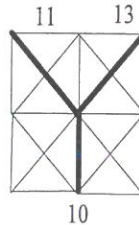


Figura 16. Reprezentarea simbolului U din alfabetul A_G

Tabelul 19. Valorile tabelare ale simbolului U din alfabetul grecesc

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|----------|------------|------------|--------|--------------|---------------------------|
| r_{10} | 10 | 1 | 0,034 | 100 | 34 |
| r_{11} | 11 | 1 | 0,034 | 121 | 45,254 |
| r_{13} | 13 | 1 | 0,0408 | 169 | 89,63 |

Pentru simbolul I, se obține $H'(U) = 168,89$.

Pentru simbolul I, se obțin datele din tabelul 20.

Tabelul 20. Valorile tabelare ale simbolului I din alfabet

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|----------|------------|------------|--------|--------------|---------------------------|
| r_9 | 9 | 1 | 0,0272 | 81 | 19,82 |
| r_{10} | 10 | 1 | 0,034 | 100 | 34 |

Pentru valorile din tabelul 20, se obține $H'(H) = 53,82$.

Tabelul 21. Valorile tabelare pentru segmentele comune simbolurilor U și I

| | α_i | δ_j | p_j | α_i^2 | $\delta_j p_j \alpha_j^2$ |
|----------|------------|------------|-------|--------------|---------------------------|
| r_{10} | 10 | 1 | 0,034 | 100 | 34 |

Pentru valorile tabelare, se obține $\theta = 34$.

Indicatorul de ortogonalitate pentru simbolurile U și I se calculează după relația:

$$H'(U, I) = \frac{\theta}{\max\{H'(U), H'(I)\}}$$

unde θ se obține asemănător cu $H'(U)$, însă doar pentru segmentele comune celor două simboluri. Se obține $H'(U, I) = 0,20$.

Aplicând relația aferentă indicatorului H' pentru toate simbolurile alfabetului, se obține matricea $M'(A_G)$. Folosind valorile obținute, se completează tabelul valorilor distincte și ale frecvențelor de apariție ale acestora.

În matricea $M'(A_G)$, se observă că a crescut diversitatea valorilor indicatorilor de ortogonalitate. Astfel, frecvențele de apariție se redistribuie între noile valori. Aceasta arată că, pentru indicatorul H' , caracterul necompensatoriu s-a îmbunătățit.

5. Analiză comparată a ortogonalității alfabetelor

Metricile de ortogonalitate evidențiază diferențele dintre entități.

Se consideră entitățile E_i și E_j . În tabelul 22, se evidențiază prezența/absența caracteristicilor în raport cu care sunt efectuate analize ale entităților E_i și E_j .

Tabelul 22. Caracteristicile entităților E_i și E_j

| Caracteristica | E_i | E_j |
|----------------|---------------|---------------|
| C_1 | γ_{1i} | γ_{1j} |
| C_2 | γ_{2i} | γ_{2j} |
| ... | ... | ... |
| C_k | γ_{ki} | γ_{kj} |
| ... | ... | ... |
| C_m | γ_{mi} | γ_{mj} |

unde γ_{kl} ia una din valorile:

- 0, dacă entitatea E_i prezintă caracteristica C_k ;
- 1, dacă entitatea E_i nu are caracteristica C_k .

Pe baza valorilor din tabelul 22, se definește indicatorul de ortogonalitate:

$$H_c(E_i, E_j) = \frac{\sum_{i=1}^m (\bar{\gamma}_{ki} \cap \gamma_{kj}) \cup (\gamma_{ki} \cap \bar{\gamma}_{kj})}{\sum_{i=1}^m (\gamma_{ki} \cup \gamma_{kj})}$$

Din alfabetul cifrelor romane, se extrag simbolurile L și C, a căror reprezentare în sistemul de referință considerat este dată în tabelul 23.

Tabelul 23. Reprezentarea simbolurilor L și C

| R | L | C | $(\bar{\gamma}_{ki} \cap \gamma_{kj}) \cup (\gamma_{ki} \cap \bar{\gamma}_{kj})$ | $\gamma_{ki} \cup \gamma_{kj}$ |
|--------------|---|---|--|--------------------------------|
| r_1 | 1 | 1 | 0 | 1 |
| r_2 | 1 | 1 | 0 | 1 |
| r_3 | 0 | 1 | 1 | 1 |
| r_4 | 0 | 1 | 1 | 1 |
| r_5 | 0 | 0 | 0 | 0 |
| r_6 | 0 | 0 | 0 | 0 |
| r_7 | 1 | 1 | 0 | 1 |
| r_8 | 1 | 1 | 0 | 1 |
| r_9 | 0 | 0 | 0 | 0 |
| r_{10} | 0 | 0 | 0 | 0 |
| r_{11} | 0 | 0 | 0 | 0 |
| r_{12} | 0 | 0 | 0 | 0 |
| r_{13} | 0 | 0 | 0 | 0 |
| r_{14} | 0 | 0 | 0 | 0 |
| r_{15} | 0 | 0 | 0 | 0 |
| r_{16} | 0 | 0 | 0 | 0 |
| r_{17} | 0 | 0 | 0 | 0 |
| r_{18} | 0 | 0 | 0 | 0 |
| r_{19} | 0 | 0 | 0 | 0 |
| r_{20} | 0 | 0 | 0 | 0 |
| Total | | | 2 | 6 |

Pentru simbolurile L și C, valoarea indicatorului de ortogonalitate este $H_c(L, C) = 0,33$.

Din alfabetele slav, respectiv grec, se preiau reprezentările caracterelor echivalente A și B.

Tabelul 24. Reprezentarea perechii (A,B) în alfabetele slav și grec

| R | Alfabet slav | | Alfabet grec | |
|-----------------|--------------|---|--------------|---|
| | A | B | A | B |
| r ₁ | 1 | 1 | 1 | 1 |
| r ₂ | 1 | 1 | 0 | 1 |
| r ₃ | 1 | 1 | 0 | 1 |
| r ₄ | 1 | 1 | 0 | 1 |
| r ₅ | 1 | 0 | 0 | 0 |
| r ₆ | 1 | 1 | 1 | 1 |
| r ₇ | 0 | 1 | 0 | 1 |
| r ₈ | 0 | 1 | 0 | 1 |
| r ₉ | 0 | 0 | 0 | 0 |
| r ₁₀ | 0 | 0 | 0 | 0 |
| r ₁₁ | 0 | 0 | 0 | 0 |
| r ₁₂ | 0 | 0 | 0 | 0 |
| r ₁₃ | 0 | 0 | 0 | 0 |
| r ₁₄ | 0 | 0 | 0 | 0 |
| r ₁₅ | 0 | 0 | 1 | 0 |
| r ₁₆ | 0 | 0 | 0 | 0 |
| r ₁₇ | 0 | 0 | 0 | 0 |
| r ₁₈ | 0 | 0 | 1 | 0 |
| r ₁₉ | 1 | 1 | 1 | 1 |
| r ₂₀ | 1 | 1 | 1 | 1 |

Se calculează indicatorii H_C și H' pentru perechile (A, B) din fiecare alfabet. Valorile indicatorilor de ortogonalitate sunt prezentate în tabelul 25.

Tabelul 25. Valorile indicatorilor H_C și H' pentru perechea (A,B)

| | Alfabet slav | Alfabet grec |
|-----------------------|--------------|--------------|
| H _C (A, B) | 0,30 | 0,63 |
| H'(A, B) | 0,96 | 0,78 |

Indicatorii H_C și H' evidențiază ortogonalitatea internă alfabetului. Diferențele între două alfabete pentru reprezentarea caracterelor similare sunt măsurate prin intermediul metricilor H_R și H_{R'} cu următoarele expresii analitice:

$$H_{R}(a_k, a_l) = \frac{H_C^i(a_k, a_l)}{H_C^j(a_k, a_l)}, \quad H_{R'}(a_k, a_l) = \frac{H_i'(a_k, a_l)}{H_j'(a_k, a_l)}$$

unde:

- $H_C^i(a_k, a_l)$ reprezintă valoarea indicatorului H_C(a_k, a_l) pentru alfabetul i;
- $H_i'(a_k, a_l)$ este valoarea indicatorului H'(a_k, a_l) pentru alfabetul i.

Pentru exemplul considerat, H_R(A, B) = 0,47 și H_{R'}(A, B) = 1,23. Astfel, valoarea indicatorului H_R(A,B) arată că, în alfabetul slav, există de două ori mai multe elemente din sistemul de referință considerat, care se refolosec la reprezentarea caracterelor A și B. Valoarea indicatorului H_{R'}(A, B) indică faptul că alfabetul grec se caracterizează de o mai mare diversitate a reprezentării.

Valorile metricilor H(A_L), H(A_S) și H(A_G) sunt evidențiate în tabelul 26.

Tabelul 26. Valorile indicatorilor de ortogonalitate ale alfabetelor

| | Alfabetul cifrelor romane | Alfabetul slav | Alfabetul grec |
|---|---------------------------|----------------|----------------|
| Ortogonalitate internă (H(A _L), H(A _S), H(A _G)) | 0,87 | 0,66 | 0,75 |

În tabelul anterior, se observă că reprezentarea cifrelor romane în sistemul de referință considerat are cel mai ridicat nivel de ortogonalitate. Astfel, diferențele în reprezentarea cifrelor romane sunt mai mari

decât în cazul alfabetelor slav, respectiv grec. O diversitate mai mare de componente ale sistemului de referință este utilizat în cazul cifrelor romane.

Diferențele de ortogonalitate internă sunt evidențiate în tabelul următor:

Tabelul 27. Valorile indicatorului H_R pentru cele 3 alfabete

| | Alfabetul cifrelor romane | Alfabetul slav | Alfabetul grec |
|---------------------------|---------------------------|----------------|----------------|
| Alfabetul cifrelor romane | - | 1,31 | 1,16 |
| Alfabetul slav | | - | 0,88 |
| Alfabetul grec | | | - |

Raportat la alfabetul cifrelor romane, indicatorul de ortogonalitate are valoarea cea mai mare pentru alfabetul slav. Aceasta înseamnă că reprezentarea caracterelor pentru alfabetul slav în sistemul de referință considerat presupune cea mai largă reutilizare a segmentelor din cele trei alfabete considerate. Acest lucru este pus în evidență și de valoarea indicatorului H_R pentru alfabetele slav și grec.

6. Concluzii

Pentru sistemul de referință considerat, determinarea metricilor de ortogonalitate evidențiază diferențele dintre simbolurile dintr-un alfabet.

Analiza ortogonalității permite proiectarea de alfabete care maximizează ortogonalitatea. Creșterea ortogonalității reprezentării caracterelor este importantă prin prisma construirii de alfabete a căror simboluri sunt bine diferențiate între ele. Acest lucru conduce la reducerea posibilității de a se face confuzii între reprezentări.

Pe de altă parte, reducerea ortogonalității alfabetelor permite simplificarea reprezentării și a sistemelor de referință. De asemenea, prin reutilizarea componentelor sistemului de referință se obțin fonturi noi plecând de la cele existente.

Bibliografie

1. IVAN, I., C. BOJA: Metode statistice în analiza software, Editura ASE, București, 2004.
2. ISAIC-MANIU, AL., C. MITRUȚ, V. VOINEAGU: Statistica pentru managementul afacerilor, Editura Economică, București, 1999.
3. IVAN, I., S. NICULESCU, C. BOJA: Clonarea bazelor de date, Revista Română de Informatică și Automatică, vol. 12, nr. 4, 2002, pp. 46 – 53.
4. IVAN, I., M. POPA, S. CAPISIZU, B. FLORESCU, L. IVAN: Characteristics of the Informatics Clones. În: Master of International Business Informatics Handbook, Editura ASE, București, 2003, pp. 207 – 223.
5. IVAN, I., P. POCATILU, M. POPA, M. SACALĂ: Information Cloning. În: Proc. of the International Symposium Regional Problems in the Context of Globalization Process, Chișinău, Republica Moldova, 9 – 10 octombrie 2002, pp. 371 – 375.
6. IVAN, I., M. POPA, I. RĂDULESCU: Tehnici și metode de agregare a textelor structurate. În: A IX-a Sesiune de Comunicări Științifice cu Participare Internațională „Știința și învățământul – fundamente ale secolului al XXI-lea”, Academia Forțelor Terestre Sibiu, 25 – 26 noiembrie 2004, ISBN 973-7809-04-1, 973-7809-03-3.
7. IVAN, I., M. POPA, S. CAPISIZU, L. BREDĂ, B. FLORESCU: Clonarea informatică, Editura ASE, București, 2003.
8. IVAN, I., M. POPA, C. TOMA, C. BOJA: Data Metrics Properties. În: Proc. of the International Symposium „Innovative Applications of Information Technologies in Business and Management”, Iași, 22-23 octombrie 2004, pp. 45 – 56.
9. IVAN, I., A. VIȘOIU, M. POPA: Ortogonalitatea – caracteristică a calității bazei de modele economice. În: Revista Română de Informatică și Automatică, vol. 14, nr. 3, 2004, pp. 89 – 100.

10. **IVAN, I., GHE. NOȘCA, S. TCACIUC, O. PÂRLOG, R. CĂCIULĂ:** Calitatea datelor, Editura INFOREC, București, 1999.
11. **IVAN, I., M. POPA, C. TOMOZEI:** Reingineria entităților text. În: Revista Română de Informatică și Automatică, vol. 15, nr. 2, 2005, pp. 15-28.
12. **NOȘCA, GHE., I. IVAN, M. POPA:** A Text Fingerprints-Based Analysis Algorithm. În: Proc. of „The 5th Biennial International Symposium SIMPEC 2004”, Brașov, 14 – 15 mai 2004, pp. 281 – 286.
13. **IVAN, I., M. POPA:** Evaluarea ortogonalității proiectelor. În: Lucrările Simpozionului Internațional România și Republica Moldova „Potențialul competitiv al economiilor naționale ale României și Republicii Moldova și posibilitățile de valorificare a lui pe piața internă, europeană și mondială”, Academia Română, Institutul Național de Cercetări Economice, Pitești, 23-25 octombrie 2003, Editura EXPERT, vol. 1, pp. 376 – 385.
14. **POPA, M.:** Structured Text Entities Metrics. În: Proc. of „The 36th Informational Scientific Symposium of METRA”, București, 26 - 27 mai 2005, pp. 486 – 491.
15. **IVAN, I., M. POPA, C. TOMA, I. RĂDULESCU:** The Aggregation of the Data Orthogonality Metrics. În: Proc. of „The 35th International Scientific Symposium of METRA”, vol. 1, București, 27 – 28 mai 2004, pp. 590 – 595.
16. **IVAN, I., M. POPA:** Ortogonalitatea produselor program orientate obiect. În: Informatica Economică, vol. 8, nr. 4, 2004, pp. 93 – 96.
17. **POPA, M.:** Sistemul caracteristicilor de calitate a datelor, ASE, București, 2005, referat doctorat.
18. **POPA, M.:** Metricile de date și proprietățile lor, ASE, București, 2005, referat de doctorat.
19. **WANG, R., M. REDDY, H. KON:** Toward Quality Data: An Attribute-based Approach. În: Decision Support Systems, nr. 13, 1995, pp. 349 – 372.
20. www.hci.com.au.
21. www.research.ibm.com.