ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XI, 3 (1970)

L. ZUBRZYCKA (Wrocław)

O WYZNACZANIU SYSTEMÓW STENOGRAFICZNYCH

1. Wstęp. Stenografia jest sztuką szybkiego pisania za pomocą umownych znaków zastępujących litery i znaczników zastępujących wyrazy i całe zwroty.

W zależności od kształtu znaków i znaczników odróżniamy systemy stenograficzne *geometryczne* i *graficzne*. Zasadnicze znaki stenografii geometrycznej zbudowane są z najkrótszych linii: z prostych kresek różnej długości, pisanych w czterech różnych kierunkach, z koła, z mniej lub więcej rozchylonego półkola oraz z punktu. Znaki te łączy się w jeden wyraz w najkrótszy sposób lub nie łączy się ich wcale, pisząc jeden znak blisko drugiego. W stenografii graficznej znaki powstałe z cząstek pisma zwykłego pisane są w kierunku poziomym lub pod kątem mniejszym od 90°. Znaki stenografii graficznej łączy się ze sobą bezpośrednio lub za pomocą małej kreseczki.

Stenografię geometryczną wprowadzili Anglicy i Francuzi, graficzną — Niemcy. Wszystkie inne kraje przejęły zasady i formy swoich systemów stenograficznych od tych trzech narodów. Dopiero w najnowszych czasach zaczęły tworzyć się systemy niezależne od tych wzorów, lepiej odpowiadające danemu językowi [6].

Spośród wielu systemów stenograficznych zarejestrowanych w USA, dwa są najbardziej rozpowszechnione i konkurują ze sobą: fonetyczna stenografia Pitmana (Pitman 1818-1897) i jej wersja zmodyfikowana przez Catona (Scientific Shorthand, 1915) oraz — również fonetyczna — kursywna stenografia Gregga. Pierwsza przeważa w krajach mówiących po angielsku poza USA, druga zaś w USA. Na uwagę zasługuje jakoby jeszcze ciągle używana w parlamencie angielskim stara już stenografia Gurneya i nowa stenografia Thomasa [11].

Stenografia Pitmana ma alfabet złożony z kresek i kawałków kół, a pogrubienie daje inną literę; samogłoski oznacza się dodatkowo za pomocą różnych akcentów jedynie wtedy, gdy istnieje obawa błędnego odczytania. Oto główne znaki alfabetu (strzałki oznaczają kierunek pisma):

Alfabet stenograficzny Pitmana

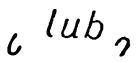
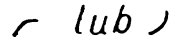
Nr	Litera	Znak stenograficzny	Nr	Litera	Znak stenograficzny
1	*P*		15	*K*	
2	*B*		16	*G*	
3	*T*		17	*M*	
4	*D*		18	*N*	
5	*CH*		19	*NG*	
6	*J*		20	*L*	
7	*F*		21	*W*	
8	*V*		22	*Y*	
9	*th*		23	*R*	
10	*TH*		24	*I*	
11	*S*		25	*O*	
12	*Z*		26	*A*	
13	*SH*		27	*E*	
14	*ZH*				

Jest kilka znaków specjalnych. Np. małe kółko na końcu wyrazu oznacza liczbę mnogą (tyle co „s"), średnie — końcówkę „ses", duże zaś „str".

Stenografia Pitmana, mimo że pisana nieciągle i kanciasto, jest szybka. Caton ulepszył jeszcze oznaczanie samogłosek.

Stenografia Gregga jest pisana ciągle, bez pogrubień i więcej okrągło. Oto podstawowe znaki:

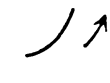
Alfabet stenograficzny Gregga

Nr	Litera	Znak stenograficzny	Nr	Litera	Znak stenograficzny
1	*K*		15	*J*	
2	*G*		16	*S*	
3	*R*		17	*SH*	
4	*L*		18	*H*	
5	*N*		19	*NG*	
6	*M*		20	*NK*	
7	*T*		21	*ă*	
8	*D*		22	*ä*	
9	*TH*		23	*ā*	
10	*P*		24	*ĭ*	
11	*B*		25	*ĕ*	
12	*F*		26	*ē*	
13	*V*		27	*ŏ*	
14	*CH*		28	*ŭ*	

Jest dużo znaków specjalnych: ted, ded, det, men, mem ten, den itd. Thomas [11] twierdzi, że opracował stenografię uwzględniającą połączenie biegłości znaków z częstością ich występowania. Ma ona dość prosty alfabet i mało dodatkowych reguł.

Niektóre znaki zmieniają sens wraz z wielkością, co stwarza możliwość nieporozumień, szczególnie przy szybkim pisaniu. W ogóle Thomas

Alfabet stenograficzny Thomasa

Nr	Litera	Znak stenograficzny	Nr	Litera	Znak stenograficzny
1	*A*		14	*N*	
2	*B*		15	*O*	
3	*C*		16	*P*	
4	*D*		17	*Q*	
5	*E*		18	*R*	
6	*F*		19	*S*	
7	*G*		20	*T*	
8	*H*		21	*U*	
9	*I*		22	*V*	
10	*J*		23	*W*	
11	*K*		24	*X*	
12	*L*		25	*Y*	
13	*M*		26	*Z*	

jest zdania, że do szybkiego pisania stenografie „kanciaste" i „nieciągłe" są lepsze, bo czytelniejsze.

W Polsce jest w użyciu kilka systemów stenograficznych [6]. Najpopularniejszy jest system graficzny Polińskiego (1861), będący adaptacją niemieckiego systemu Gabelsbergera. Zdaniem stenografów nie odpowiada on naszemu językowi i piszący nim nie osiągają wielkich szybkości (rekord polski wynosi 300-360 zgłosek na minutę, gdy tymczasem stenografowie czescy, piszący swoim własnym systemem, przekroczyli już 500 zgłosek na minutę ([4] i [9]). Najnowszy system JSSP (Jednolity System Stenografii Polskiej, 1958) jest adaptacją radzieckiego syste-

mu Sokołowa, który z kolei jest oparty na niemieckim systemie Scheithausera i Braunsa [9]. I ten system nie wnosi żadnych nowych ułatwień.

A oto główne zarzuty stawiane JSSP przez wybitnych praktyków i nauczycieli:

1. nie uwzględnia on w dostatecznym stopniu właściwości języka polskiego;
2. nie odznacza się prostotą i jasnością teorii;
3. duże podobieństwo znaków utrudnia czytelność pisma.

Istnieje więc potrzeba opracowania nowego systemu dostosowanego do języka polskiego. Ten nowy system powinien łączyć szybkość pisania znaku z częstością jego występowania w naszym języku, a zatem musi być oparty o statystyczne badania struktury języka polskiego.

W jaki sposób statystyka elementów języka ma być użyta przy wyznaczaniu systemu stenograficznego? Na podstawie znajomości istniejących systemów stenograficznych można wyróżnić trzy takie sposoby. Pierwszy — to wykorzystanie statystyki liter do tego, żeby częściej występującym literom przydzielić szybsze w pisaniu (bieglejsze) znaki alfabetu stenograficznego. Drugi — to wykorzystanie statystyki grup spółgłoskowych do tego, żeby częściej występującym grupom przydzielić bieglejsze znaki spoza alfabetu stenograficznego. Ograniczenie się do grup spółgłoskowych, rozumianych tu jako grupy spółgłosek oddzielone samogłoskami, usprawiedliwione jest tym, że w niektórych istniejących systemach stenograficznych przyjęta jest zasada pisania spółgłosek i symbolizowania samogłosek za pomocą odpowiedniego łączenia grup spółgłoskowych [6]. Trzeci — to użycie statystyki wyrazów do wyłowienia często występujących wyrazów i ewentualnie zwrotów, ażeby zastąpić je specjalnymi skrótami. O takim wykorzystaniu statystyki mówi np. K. Szuter w swym referacie (zob. [10]).

W punkcie 6 tej pracy zostanie przedstawiona metoda wyznaczania optymalnego systemu stenograficznego, gdy znany jest czas pisania poszczególnych znaków stenograficznych i częstości elementów języka, którym te znaki mają być przyporządkowane. Zanim to uczynimy, najpierw musimy się zastanowić nad tym, jak wyznaczyć częstości elementów języka polskiego, takich jak litery, grupy spółgłoskowe, itd. Zadania te omówimy w punktach 2, 3 i 4. Na końcu pracy przedstawimy projekt nowego alfabetu stenograficznego, wyznaczonego na podstawie próbki języka polskiego liczącej 4000 wyrazów.

2. Zagadnienie pobierania próbek przy badaniu częstości elementów języka. Przy jakimkolwiek badaniu częstości występowania liter, wyrazów, czy innych elementów języka trzeba rozważyć pewne podstawowe zagadnienia dotyczące pobierania próbek:

1. Ile wyrazów należy wziąć do próbki?
2. Jak należy pobrać próbkę wyrazów z języka?

Na tak postawione pytania nie można jednak dać odpowiedzi z tych samych powodów, z jakich nie można odpowiedzieć na pytanie, ile osób należy pomierzyć, aby wyznaczyć średni wzrost ludzi w jakiejś populacji. Aby móc odpowiedzieć na interesujące nas pytanie, trzeba je w odpowiedni sposób sprecyzować. W przypadku średniego wzrostu, to sprecyzowanie pytania uzyskuje się np. w ten sposób, że pytamy: ile osób należy pomierzyć, ażeby oszacować średni wzrost z błędem standardowym nie przekraczającym 1/2 cm. I to jeszcze nie wystarczy do wyznaczenia potrzebnej liczby osób. Przy takich samych wymaganiach co do standardowego błędu oszacowania średniej potrzebna będzie tym większa ilość osób do próbki, im bardziej różnią się między sobą co do wzrostu osoby badanej populacji. Trzeba więc znać ponadto jakąś charakterystykę populacji ze względu na rozmaitość wzrostu poszczególnych osób. Charakterystykę taką uzyskuje się zazwyczaj z niezbyt licznej próbki orientacyjnej. I w naszym wypadku musimy więc zrobić dwie rzeczy. Odpowiednio sprecyzować pytanie i ze stosownie pobranej próbki orientacyjnej wyznaczyć te charakterystyki wchodzącej w rachubę masy językowej, która będzie grała rolę populacji w naszych rozważaniach.

W języku polskim mamy 9 samogłosek i 23 spółgłoski. Z 23 spółgłosek można teoretycznie utworzyć $23^2 = 529$ dwuliterowych grup spółgłoskowych, oraz $23^3 = 12167$ grup trzyliterowych. Oczywiście, w języku polskim nie spotkamy wszystkich grup spółgłoskowych wynikających z dowolnego komponowania liter ze sobą i wystąpią tylko niektóre z możliwych. Od tego, ile różnych grup spółgłoskowych w języku polskim naprawdę występuje i jak bardzo różnią się co do częstości te, które występują, zależy wielkość masy lingwistycznej, którą trzeba opracować przy statystyce grup spółgłoskowych. Powiedzieliśmy, że rolę populacji w naszych rozważaniach będzie odgrywała masa językowa, to jest ten język, który mamy badać. Powiedziałby ktoś, że język polski, to to wszystko, co zostało lub będzie po polsku napisane i powiedziane. My jednak musimy z tego wszystkiego wyodrębnić jakąś populację w sensie statystycznym, by móc do niej stosować metody statystyczne. Dobrą populacją jest np. książka. Wiadomo jednak, że pisać można na różne tematy i że w zależności od tematu i autora będzie używany coraz to inny zasób słów, komponowane będą zadnia w coraz to inny sposób, że więc dzieła poświęcone różnym tematom będą różniły się między sobą tak co do słownictwa, jak i co do stylu. Dlatego szukając częstych słów, dla których warto w systemie stenograficznym stworzyć specjalne skróty, konieczne jest wyodrębnienie tego stenograficznego języka polskiego. To jest pierwsze i podstawowe zadanie, które musi być rozwiązane na

samym początku. Bez zdefiniowania języka stenograficznego, jako populacji w sensie statystycznym, nie można w rozsądny sposób mówić o pobieraniu prób i temu podobnych rzeczach.

O ile jednak nikt nie wątpi, że powyższe uwagi mają pierwszorzędne znaczenie dla statystyki słownictwa i zwrotów, to można się spodziewać, iż dla statystyki liter i grup spółgłoskowych dobór tekstów nie będzie odgrywał wielkiej roli. Mianowicie, można się spodziewać, że różne teksty polskie, różniące się tematyką, słownictwem i stylem, będą miały jednak jednakową, charakterystyczną dla języka polskiego strukturę grup spółgłoskowych. Aby się o tym przekonać, porównaliśmy co do częstości występowania liter, grup spółgłoskowych i wyrazów cztery teksty, po 1000 wyrazów każdy, jak najbardziej różne pod względem treści.

Opis tego porównania podano w punkcie 3.

3. Czy teksty różnią się pod względem częstości liter, grup spółgłoskowych i wyrazów? Do badania wzięliśmy cztery teksty: polityczno-gospodarczy, naukowo-historyczny, handlowy i literacki.

Obecnie przystępujemy do opisu tego porównania.

a. *Porównanie częstości liter.* Z próby zostały usunięte cyfry rzymskie i arabskie, nazwiska i nazwy miast. Skróty były czytane jako pełne słowa. Pierwsze porównanie, jakie zrobiliśmy, to było porównanie tych czterech tekstów co do częstości liter. Zgodność frekwencji w poszczególnych tekstach zbadaliśmy za pomocą testu χ^2. Uwagi metodyczne na temat zastosowania testu χ^2 podane są w punkcie 5. Jako wartość testu z 93 stopniami swobody otrzymaliśmy liczbę

$$\chi^2_{93} = 294{,}69.$$

Jak wiadomo ([2], str. 243), jeśli χ^2 jest zmienną losową o rozkładzie chi-kwadrat z n stopniami swobody, to zmienna losowa $\sqrt{2\chi^3_n} - \sqrt{2n-1}$ ma rozkład prawdopodobieństwa, który przy n rosnącym do ∞ zmierza do rozkładu normalnego z wartością oczekiwaną 0 i dyspersją 1.

Do wzoru $\sqrt{2\chi^2_n} - \sqrt{2n-1}$ podstawiliśmy $\chi^2_n = 294{,}69$ oraz $n = 93$ i otrzymaliśmy liczbę 10,67, co oznacza, że różnice między tekstami, jeśli idzie o częstość występowania poszczególnych liter, są bardzo istotne.

Oznacza to, że różnice w częstościach liter są na tyle wyraźne, że na ich podstawie można by się pokusić o orzeczenie, z jakiego tekstu pochodzi próbka, chociaż litery częste w jednym tekście są na ogół częste też i w trzech pozostałych, a litery rzadko występujące w jednym tekście rzadko występują i w innych (tablica 1).

TABLICA 1. Frekwencje liter w poszczególnych tekstach

Litera	Tekst				
	I	II	III	IV	Razem
A	502	542	641	474	2159
Ą	70	69	79	57	275
B	90	53	80	63	286
C	234	257	241	226	958
Ć	30	17	9	14	70
D	219	212	269	163	863
E	483	459	443	384	1769
Ę	76	81	59	81	297
F	25	18	16	8	67
G	93	67	84	86	330
H	54	60	46	86	246
I	485	589	423	449	1946
J	101	148	115	114	478
K	204	247	247	212	910
L	109	147	97	113	466
Ł	114	149	130	149	542
M	162	158	144	167	631
N	380	316	367	293	1356
Ń	1	21	17	15	54
O	452	480	516	398	1846
Ó	58	60	67	42	227
P	205	179	193	159	736
R	271	249	341	232	1093
S	276	268	277	217	1038
Ś	40	39	30	48	157
T	248	203	229	200	880
U	184	163	185	152	684
W	327	304	368	227	1226
Y	301	209	297	251	1058
Z	356	378	431	310	1475
Ż	46	47	49	46	188
Ź	14	8	5	3	30
Razem	6210	6197	6495	5439	24341

b. *Porównanie częstości grup spółgłoskowych.* W naszym materiale zaobserwowaliśmy 312 różnych grup spółgłoskowych: 22 grupy 1-literowe, 152 grupy 2-literowe, 122 grupy 3-literowe, 13 grup 4-literowych i 3 grupy 5-literowe. Zgodność frekwencji grup spółgłoskowych w badanych tekstach obliczaliśmy za pomocą testu χ^2. Badania przeprowadziliśmy dla grup spółgłoskowych, które wystąpiły 20 razy i więcej w czterech tekstach razem (tablica 2). Jako wartość testu χ^2 z 162 stopniami swobody otrzymaliśmy liczbę $\chi^2_{162} = 688{,}44$, a jako wartość zmiennej losowej $\sqrt{2\chi^2_n} - \sqrt{2n-1}$ liczbę 19,14.

TABLICA 2. Frekwencje grup spółgłoskowych w poszczególnych tekstach

Grupa spółgłoskowa	Tekst				Razem
	I	II	III	IV	
B	70	36	44	43	193
C	76	96	92	68	332
Ć	21	13	4	11	49
D	80	91	144	75	390
F	21	15	16	7	59
G	57	36	53	56	202
J	94	138	93	95	420
K	77	114	131	85	407
L	54	69	62	63	248
Ł	54	112	72	99	337
M	139	130	124	131	524
N	202	181	223	169	775
Ń	—	7	9	7	23
P	63	93	76	89	321
R	77	88	99	57	321
S	62	79	72	68	281
T	85	83	100	90	358
W	234	191	283	145	853
Z	104	82	124	74	384
Ż	38	31	36	40	145
BR	2	7	22	2	33
CH	49	50	38	64	201
CZ	22	31	34	21	108
DN	13	8	29	9	59
DR	2	5	9	8	24
DS	15	1	5	—	21
DZ	44	64	32	22	162
GŁ	5	5	6	13	29
GR	16	9	13	2	40
KŁ	5	3	15	7	30
KR	4	13	6	13	36
KT	21	10	15	12	58
LK	7	8	2	8	25
LN	13	10	9	3	35
NK	8	4	4	6	22
NN	12	11	5	4	32
NT	16	8	5	4	33
PŁ	33	1	2	7	43
PR	20	19	36	3	78
RZ	14	11	40	13	78
SK	41	10	6	14	71
SŁ	3	9	9	4	25
SP	13	7	9	7	36
ST	48	43	49	23	163
SZ	13	12	44	26	95
ŚĆ	20	12	15	11	58
WN	18	10	21	13	62

TABLICA 2 c.d

Grupa spółgłoskowa	Tekst I	II	III	IV	Razem
ZM	7	12	2	4	25
ZN	7	8	12	6	33
ZW	7	17	12	8	44
CZN	21	16	11	4	52
PRZ	51	31	31	36	149
SPR	2	3	15	—	20
RSTW	14	2	4	—	20
Inne grupy	300	357	299	366	1322
Razem	2494	2512	2723	2215	9944

Różnice są bardzo istotne. Dowodzi to, wbrew naszym pierwotnym zbyt optymistycznym przypuszczeniom, że dla ustalenia częstości grup spółgłoskowych nie jest obojętne, na podstawie jakiego tekstu to się robi. Konieczne jest przeto opracowywanie takich tekstów, jakie się stenografuje.

c. *Porównanie częstości wyrazów.* Na 4000 zbadanych wyrazów występowanie w poszczególnych tekstach różnych wyrazów przedstawia się następująco:

tekst	I	II	III	IV	razem
różnych wyrazów	554	657	545	695	2176

Za różne uznaliśmy wyrazy różniące się choćby jedną literą. Wyrazów, które wystąpiły tylko w jednym tekście, jest 1913.

Liczba wyrazów (n) o jednakowych częstościach (f)

n	1699	246	97	44	25	12	11	5	6	3	4	1	3
f	1	2	3	4	5	6	7	8	9	10	11	12	13

n	2	2	1	1	1	1	1	1	1	1	1	1	1	1
f	14	16	17	18	19	20	21	23	27	32	34	38	52	60

n	1	1	1	1
f	72	86	115	155

Jeżeli wszystkie formy czasownika i wszystkie przypadki rzeczownika będziemy uważać za ten sam wyraz, to występowanie różnych wyrazów w poszczególnych tekstach przedstawia się następująco:

tekst	I	II	III	IV	razem
różnych wyrazów	472	555	444	613	1690

Już z tego pobieżnego porównania widać, że teksty bardzo się różniły pod względem słownictwa.

4. Jaka masa jest potrzebna do uporządkowania elementów języka według częstości? W niniejszym punkcie podamy, w jaki sposób można z opracowanych próbek tekstu uzyskać oszacowanie masy językowej niezbędnej do uporządkowania elementów języka według częstości. Mianowicie, postawiliśmy sobie pytanie, o ile trzeba by powiększyć próbkę, aby takie różnice częstości, jakie zaobserwowaliśmy w naszej próbce, można było uważać za istotnie różne od zera. A jak się bada istotność różnic frekwencji, opiszemy na przykładzie liter *a* i *o*.

Przy badaniu hipotezy, że litery *a* oraz *o* mają takie same częstości, rozumujemy następująco: literze *a* przyporządkujmy zmienną losową ξ_a; zmienna ta przyjmuje wartość 1, gdy litera wylosowana z tekstu jest literą *a*, i wartość 0, gdy jest inną literą. Podobnie, literze *o* przyporządkujmy zmienną losową ξ_o, która przyjmuje wartość 1, gdy litera wylosowana z tekstu jest literą *o*, i wartość 0, gdy jest inną literą. O literach losowanych z tekstu zakładamy, że jest to losowanie ze zwracaniem z urny zawierającej jednakowe frakcje liter *a* i liter *o* oraz inne litery.

Gdyby ta wspólna frakcja liter *a* i *o* była p, to łączny rozkład prawdopodobieństwa zmiennych losowych ξ_a i ξ_o byłby następujący:

$$\begin{aligned} &\mathrm{P}(\xi_a = 1,\ \xi_o = 1) = 0,\\ &\mathrm{P}(\xi_a = 1,\ \xi_o = 0) = \mathrm{P}(\xi_a = 0,\ \xi_o = 1) = p,\\ &\mathrm{P}(\xi_a = 0,\ \xi_o = 0) = 1-2p. \end{aligned}$$

Każda litera wylosowana z tekstu daje nam wartości pary zmiennych losowych ξ_a i ξ_o; traktujemy je jako niezależne obserwacje tej pary zmiennych losowych.

Niech więc $\xi_a^{(i)}$ oznacza wartość zmiennej ξ_a dla i-tej litery z naszej próbki; podobne znaczenie nadajemy symbolowi $\xi_o^{(i)}$ w odniesieniu do litery *o*. Wówczas $\bar{x} = (1/n)\sum_i \xi_a^{(i)}$, gdzie n jest liczbą liter w próbce, jest zaobserwowaną frekwencją litery *a*, a $\bar{y} = (1/n)\sum_i \xi_o^{(i)}$ jest zaobserwowaną frekwencją litery *o*.

Pytamy, czy częstości liter *a* i *o* są różne, albo inaczej, czy różnica $\bar{x}-\bar{y}$ jest istotnie różna od zera. Możemy napisać, że

$$\bar{x}-\bar{y} = (1/n)\sum_i (\xi_a^{(i)} - \xi_o^{(i)}),$$

a więc różnica, o którą chodzi, daje się przedstawić jako średnia arytmetyczna niezależnych zmiennych losowych. Aby ocenić istotność tej różnicy, skorzystamy z tego, że jako średnia wielu zmiennych losowych o jednakowym rozkładzie ma ona na mocy centralnego twierdzenia rachunku prawdopodobieństwa rozkład normalny. Dla oceny istotności pozostaje wobec tego podzielić tę różnicę przez jej dyspersję $D(\bar{x}-\bar{y})$.

Wobec przyjętych założeń

$$D^2(\bar{x}-\bar{y}) = (1/n)D^2(\xi_a-\xi_o).$$

Z łącznego rozkładu prawdopodobieństwa zmiennych ξ_a i ξ_o znajdujemy $D^2(\xi_a-\xi_o) = 2p$, wobec tego

$$D^2(\bar{x}-\bar{y}) = 2p/n.$$

Przy badaniu istotności za p podstawialiśmy $\frac{1}{2}(\bar{x}+\bar{y})$, a więc liczbę

$$t = \frac{\bar{x}-\bar{y}}{D(\bar{x}-\bar{y})} = \frac{(\bar{x}-\bar{y})\sqrt{n}}{\sqrt{2p}} = \frac{\sqrt{n}(\bar{x}-\bar{y})}{\sqrt{\bar{x}+\bar{y}}} =$$

$$= \frac{n(\bar{x}-\bar{y})}{\sqrt{n(\bar{x}+\bar{y})}} = \frac{\sum \xi_a^{(i)} - \sum \xi_o^{(i)}}{\sqrt{\sum \xi_a^{(i)} + \sum \xi_o^{(i)}}}$$

traktowaliśmy jako obserwacje zmiennej losowej o rozkładzie normalnym o wartości oczekiwanej 0 i dyspersji 1. Za istotnie różne od zera uważaliśmy różnice, dla których $|t| > 3{,}29$, co odpowiada mniej więcej poziomowi istotności 0,001. Badanie przeprowadziliśmy dla łącznej frekwencji z wszystkich tekstów, przy czym badaliśmy tylko istotność różnic liter, osobno samogłosek i osobno spółgłosek, sąsiednich pod względem frekwencji.

Okazało się, że 6 różnic było istotnie rożnych od zera, a reszta, tj. 24 różnice, okazała się nieistotnie różna od zera. Nie należy uważać, że w tych wypadkach, kiedy różnice były nieistotne, odpowiednie litery mają jednakowe częstości. Ale należy przypuszczać, że litery te różnią się co do częstości na tyle mało, że próbka o takiej objętości jak nasza jest jeszcze za mała, żeby móc stwierdzić z dostateczną pewnością, która z liter jest częstsza. Żeby ocenić, jakie powiększenie próbki wystarczyłoby do stwierdzenia istotności różnicy, rozumowaliśmy w następujący sposób: gdyby próbkę powiększyć k razy, to frekwencja każdej litery z osobna zwiększyłaby się też k razy. A wobec tego liczba t zwiększyłaby się $\sqrt{k}$ razy. Uznajemy, że częstości liter są różne, gdy $|t| > 3{,}29$. Dla tych więc par liter, dla których $|t|$ okazało się $< 3{,}29$, postawiliśmy sobie pytanie, dla jakiego k mamy $|t|\sqrt{k} = 3{,}29$. Frekwencje tych liter, wartości $|t|$ i k podajemy w tablicy 3. Dla grup spółgłoskowych przeprowadzi-

TABLICA 3. Wartości $|t|$ i k dla liter

Litera	Frekwencja	$\|t\|$	k	Litera	Frekwencja	$\|t\|$	k
Samogłoski				K	910		
						0,71	21,47
A	2159			T	880		
		3,32				0,41	64,38
I	1946			D	863		
		1,62	4,12			3,18	1,07
O	1846			P	736		
		1,28	6,60			2,84	1,34
E	1769			M	631		
		13,37				2,60	1,60
Y	1058			Ł	542		
		8,96				2,00	2,71
U	684			J	478		
		12,37				0,39	71,17
Ę	297			L	466		
		0,92	12,79			4,82	
Ą	275			G	330		
		2,14	2, 36			1,77	3,46
Ó	227			B	286		
						1,73	3,62
				H	246		
Spółgłoski						2,78	1,40
				Ż	188		
						1,67	3,88
Z	1475	2,24	2,16	Ś	157		
						5,77	
N	1356	2,56	1,65	Ć	70		
						0,26	160,12
W	1226	2,76	1,42	F	67		
						1,18	7,77
R	1093	1,19	7,65	Ń	54		
						2,62	1,58
S	1038	1,79	3,38	Ź	30		
C	958	1,11	8,79				

liśmy takie same badania, jak dla liter, z ograniczeniem tylko do grup spółgłoskowych, które wystąpiły co najmniej 20 razy. Frekwencje tych grup spółgłoskowych oraz wartości $|t|$ i k podajemy w tablicy 4.

Na tle tych wyników można więc się spodziewać, że przy 25-krotnym powiększeniu próbki, już poza 3 parami liter bardzo mało różniących się co do frekwencji, uzyskalibyśmy dokładne uporządkowanie liter według częstości. Dla dokładnego uporządkowania grup spółgłoskowych potrzebne by było znacznie większe, mniej więcej 800-krotne po-

rzycka

i k dla grup spółgłoskowych

Grupa spółgłoskowa	Frekwencja	$\|t\|$	k
1	2	3	4
CZ	108		
		0,91	13,07
SZ	95		
		1,29	6,50
PR	78		
		0,08	1691,27
RZ	78		
		0,57	33,32
SK	71		
		0,78	17,79
WN	62		
		0,27	148,47
DN	59		
		0,09	1336,34
F	59		
		0,09	1336,34
KT	58		
		0,09	1336,34
ŚĆ	58		
		0,57	33,32
CZN	52		
		0,30	120,28
Ć	49		
		0,52	40,03
ZW	44		
		0,11	894,55
PŁ	43		
		0,33	99,40
GR	40		
		0,46	51,15
KR	36		
		0,12	751,69
SP	36		
		0,12	751,69
LN	35		
		0,24	187,91
NT	33		
		0,12	751,69
ZN	33		
		0,12	751,69
BR	33		
		0,12	751,69

lla obliczenia k przyjęliśmy, że różnica

TABLICA 4, c.d.

Grupa spółgłoskowa	Frekwencja	$\|t\|$	k	Grupa spółgłoskowa	Frekwencja	$\|t\|$	k
1	2	3	4	1	2	3	4
NN	32			Ń	23		
		0,25	173,19			0 15	481,06
KŁ	30			NK	22		
		0,13	640,49			0,15	481,06
GŁ	29			DS	21		
		0,54	37,11			0,16	422,84
LK	25			SPR	20		
		0,14	552,25			0,16	422,84
SŁ	25			RSTW	20		
		0,14	552,25				
ZM	25			Inne grupy o frekwencji mniejszej od 20,			
		0,14	552,25				
DR	24			razem	1322		
		0,15	481,06				

większenie próbki. A więc, dla uporządkowania liter próbka powinna wynosić 100000 wyrazów, a dla uporządkowania grup spółgłoskowych 3200000 wyrazów, jeśli oczywiście potrzebne jest aż tak dokładne uporządkowanie.

5. Uwagi krytyczne. Próbowaliśmy odpowiedzieć na dwa pytania: jak wielką próbkę wyrazów należy pobrać, żeby uporządkować elementy języka polskiego według częstości, i jak należy ją pobrać? Jeśli uporządkowanie ma być dokładne, to dla uporządkowania liter potrzeba 100000 wyrazów, a dla uporządkowania grup spółgłoskowych 3200000 wyrazów, nie licząc w tym wyrazów częstych, które zostaną zastąpione znacznikami. Przekonaliśmy się, że nie jest obojętne, na podstawie jakich tekstów oblicza się te częstości. Dlatego przede wszystkim trzeba zdefiniować język stenograficzny, następnie z niego pobrać reprezentacyjną próbkę o podanej liczebności, sporządzić na jej podstawie statystyki, a w końcu zbadać za pomocą testów statystycznych, czy uporządkowanie jest istotne.

Co to jest *język stenograficzny*? Najwłaściwsze wydaje się przyjąć, że to jest to, co się obecnie stenografuje, ewentualnie można doń jeszcze doliczyć to, o czym sądzi się, że powinno być stenografowane. Natomiast w żadnym wypadku nie są to stenogramy, ponieważ stenografowie nie stenografują dosłownie. Ten żywy język stenograficzny, który będziemy badać, musi być nagrany na taśmę magnetofonową.

Została nam jeszcze jedna sprawa do omówienia: krytyka metod zastosowanych do porównania tych czterech tekstów. Dla uproszczenia

rozumowaliśmy tak, jak gdyby rozważane elementy języka (litery, grupy spółgłoskowe) były losowane niezależnie. Dla liter zakładaliśmy, że są losowane z urny zawierającej ustalone frakcje poszczególnych liter. Dla spółgłosek zakładaliśmy, że tworząc polski tekst losujemy niezależnie na przemian z dwóch urn, z których jedna zawiera grupy spółgłoskowe, druga — samogłoski i odstępy między wyrazami, w proporcjach nie zmieniających się w czasie losowania. W jaki sposób założenie tej niezależności mogło wpłynąć na wynik? Otóż, błędy wynikłe z tych uproszczeń mogą, w zależności od prawdziwej struktury języka, przesunąć wynik w jedną lub drugą stronę. Układ *aabaabaabaabaabaa*, gdzie po każdej literze *b* następują dwie litery *a*, jest przykładem tekstu, w którym bardzo mała próbka wystarczy do ustalenia i porównania częstości liter. Inny przykład to tekst, gdzie serie liter następują po sobie, w rodzaju *aaaaabbbbbbbaaaaaaabbbaabbbbbbbbaaa*; w tekście tego rodzaju ustalenie częstości liter wymagałoby próbki znacznie większej. Prawdziwy tekst ma strukturę zawartą między tymi dwoma ekstremami. Można w nim dopatrzyć się zarówno efektu pierwszego, jak i drugiego, bo tekst złożony ze słów wyklucza serie jednakowych liter i doprowadza do ich przetasowania. Ponadto, efektu seryjnego można się dopatrywać w tym, że w sąsiednich odcinkach tekstu występują podobne wyrażenia. Z tego powodu można sądzić, że użyte przez nas założenie niezależności, które pozwala na stosowanie testu χ^2, nie wpływa zbytnio na wynik. Dokładniejsze oszacowanie omawianych efektów wymagałoby przypuszczalnie żmudnego opracowania znacznie większej próbki tekstu niż jest to potrzebne do ustalenia częstości elementów języka polskiego.

6. Jak wyznaczyć alfabet stenograficzny? W jaki sposób znajomość częstości występowania poszczególnych liter ma być wykorzystana przy wyznaczaniu alfabetu stenograficznego? Alfabet stenograficzny należy tak wyznaczyć, żeby średni czas napisania tekstu w danym języku był najkrótszy. Można powiedzieć inaczej — żeby średni czas napisania jednego znaku w danym systemie i w danym języku był minimalny. Niektóre znaki stenograficzne pisze się szybciej na początku wyrazu, a wolniej w środku, gdy się je łączy ze znakami poprzednimi. Zatem, badając szybkość pisania znaku stenograficznego, musimy to uwzględnić, wobec czego potrzebna nam jest również znajomość częstości występowania liter na początku wyrazu, w środku i na końcu.

Załóżmy, że mamy dwie tablice: w tablicy T_1 podane są dla wszystkich liter prawdopodobieństwa pojawienia się na początku wyrazu, w środku i na końcu, w tablicy T_2 — czas pisania znaku stenograficznego na początku wyrazu, w środku i na końcu.

W jaki sposób należy znaki przydzielić literom, żeby średni czas napisania jednego znaku stenograficznego był najmniejszy?

Niech literze o numerze i występującej na początku wyrazu z prawdopodobieństwem p_{i1}, w środku wyrazu — z prawdopodobieństwem p_{i2} i na końcu — z prawdopodobieństwem p_{i3} przydzielony będzie znak stenograficzny o czasie pisania na początku wyrazu t_{i1}, w środku — t_{i2} i na końcu — t_{i3}.

Wyrażenie

$$S = \sum_{i=1}^{N} \sum_{j=1}^{3} t_{ij} p_{ij},$$

gdzie N jest ilością liter alfabetu, będzie średnim czasem napisania jednego znaku w danym systemie i w danym języku. Zadaniem naszym jest tak przydzielić znaki literom, żeby wyrażenie S osiągnęło minimum.

Odwrotność tej wielkości $E = 1/S$, będziemy nazywać *efektywnością* danego systemu stenograficznego dla danego języka.

Oczywiście czas napisania danego znaku stenograficznego w środku wyrazu (w systemie, w którym wyrazy łączy się) zależy od tego, po jakim znaku pisze się dany znak. W systemach graficznych dwa znaki stenograficzne łączy się ze sobą bezpośrednio lub za pomocą kreseczki połączeniowej. Przypuśćmy, że literze g i literze t przydzielono znaki, które wymagają dopisania dodatkowej kreseczki przy łączeniu ich w kolejności gt. W tekstach polskich litery g i t w kolejności gt nie występują wcale i przy takim przydziale znaków szybkość pisania znaku przydzielonego literze t w środku wyrazu będzie inna niż szybkość tego samego znaku, gdy przydzielimy go literze d, bo kolejność gd w języku polskim jest częsta. I dlatego szybkość pisania znaków należałoby obliczać dla ustalonego alfabetu stenograficznego. Ponieważ prawdopodobieństwa pojawienia się danej litery w tekście nie dadzą się wyznaczyć dokładnie, będą obciążone dużym błędem, a poza tym będą zmieniały się w czasie razem ze zmianą języka, nie warto więc wyznaczać aż tak dokładnie czasu pisania znaków stenograficznych. Jako czas pisania znaku w środku wyrazu proponujemy przyjąć czas pisania tego znaku przy najgorszym sąsiedztwie, wymagającym dopisania dodatkowej kreseczki połączeniowej, traktując to jako wielkość stałą niezależną od przydziału znaków literom.

Niech będzie dana tablica T_1 dla języka polskiego i tablica T_2 dla danego tworzywa stenograficznego. W nagłówkach wierszy tablicy T_1 niech będą wypisane litery, a w nagłówkach kolumn topografia danej litery, tzn., czy występuje ona jako samodzielny wyraz, czy na początku wyrazu, w środku, czy na końcu. Tablica T_1 ma zatem 32 wiersze i 4 kolumny. W nagłówkach kolumn tablicy T_2 niech będą wypisane znaki stenograficzne, a w nagłówkach wierszy topografia danego znaku, czy

jest on pisany osobno, na początku wyrazu, w środku, czy na końcu. Tablica T_2 ma zatem 4 wiersze i 32 kolumny.

Pomnóżmy teraz tablicę T_1 przez tablicę T_2 zgodnie z regułą mnożenia macierzy i wynik nazwijmy macierzą T_3. Wówczas S będzie równe sumie wyrazów stojących na głównej przekątnej macierzy T_3 (*ślad macierzy*). Zadanie nasze możemy teraz sformułować tak: znaleźć takie uporządkowanie kolumn macierzy T_2 lub, co na jedno wychodzi, kolumn macierzy T_3, żeby ślad macierzy T_3 osiągnął minimum.

Różnych uporządkowań n elementów jest $n!$, ale nie musimy wypróbowywać wszystkich permutacji, żeby wybrać optymalne uporządkowanie. Wystarczy zauważyć, że przy przestawianiu kolumn cyklicznie zmieniają się tylko wyrazy przekątnej w kolumnach odpowiadających znakom z danych cykli. Różnice wynikające z cyklicznego przestawienia kolumn dla cykli rozłącznych sumują się. Jak wiadomo z algebry ([7], str. 38), każdą permutację można przedstawić za pomocą cykli rozłącznych. A wobec tego dla rozwiązania naszego zadania wystarczy zbadać wszystkie cykliczne przestawienia danych znaków i jeśli one nie polepszą efektywności, uznać dane uporządkowanie za optymalne.

Wszystkich permutacji n elementów jest $n!$, a cykli, jakie można utworzyć z tych elementów, pomijając cykle jednoelementowe, które nie zmieniają permutacji, jest

$$\sum_{k=2}^{n} \binom{n}{k} (k-1)!.$$

A wobec tego prawdziwa jest nierówność

$$\sum_{k=2}^{n} \binom{n}{k} (k-1)! < n! \quad \text{dla} \quad k \leqslant n. \tag{1}$$

Ile się zyskuje na tym, że zamiast $n!$ permutacji rozpatrujemy tylko $\sum_{k=2}^{n} \binom{n}{k} (k-1)!$ cykli?

Obliczmy stosunek ilości różnych cykli do ilości permutacji dla n elementów,

$$\frac{\sum_{k=2}^{n} \binom{n}{k} (k-1)!}{n!},$$

i utwórzmy ciąg o wyrazach

$$a_n = \frac{\sum_{k=2}^{n} \binom{n}{k} (k-1)!}{n!}$$

Można to napisać inaczej, wprowadzając $n!$ pod znak sumy, i wówczas

$$a_n = \sum_{k=2}^{n} \frac{1}{k(n-k)!}. \tag{2}$$

TWIERDZENIE. *Zachodzi następująca relacja*:

$$\lim_{n\to\infty} a_n = 0.$$

Twierdzenie to wynika bezpośrednio z następujących dwóch lematów.

LEMAT 1. *Dla wyrazów ciągu* $\{a_n\}$ *podanych wzorem* (2) *prawdziwa jest nierówność*

$$\frac{a_{n+1}}{a_n} < \frac{n}{n+1} \quad \text{dla } n \geqslant 7. \tag{3}$$

Dowód. Wypiszmy składniki sum licznika i mianownika wyrażenia po lewej stronie wzoru (3):

$$\begin{aligned} a_{n+1} &= \frac{1}{2(n-1)!} + \frac{1}{3(n-2)!} + \ldots + \frac{1}{(n-1)2!} + \frac{1}{n1!} + \frac{1}{(n+1)0!}, \\ a_n &= \frac{1}{2(n-2)!} + \frac{1}{3(n-3)!} + \ldots + \frac{1}{(n-1)1!} + \frac{1}{n0!}. \end{aligned} \tag{4}$$

W górnej sumie jest o jeden składnik więcej. Żeby te sumy miały tę samą ilość składników, dodajmy dwa pierwsze wyrazy górnej sumy do siebie i traktujmy to jako pierwszy składnik tej sumy:

$$\frac{1}{2(n-1)!} + \frac{1}{3(n-2)!} = \frac{2n+1}{6(n-1)!}. \tag{5}$$

Utwórzmy stosunki kolejnych wyrazów górnej sumy do wyrazów dolnej sumy. Otrzymamy następujący ciąg:

$$\frac{2n+1}{3(n-1)}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots, \frac{n-1}{n}, \frac{n}{n+1}. \tag{6}$$

Pomijając pierwszy wyraz, wszystkie pozostałe nie przekraczają $n/(n+1)$. Obliczmy, dla jakich n pierwszy wyraz ciągu (6) będzie mniejszy od $n/(n+1)$, czyli rozwiążmy nierówność

$$\frac{2n+1}{3(n-1)} < \frac{n}{n+1}.$$

Nierówność ta jest prawdziwa dla wszystkich $n \geqslant 7$, jeśli interesują nas tylko n naturalne. Dla $n \geqslant 7$ wszystkie wyrazy z wyjątkiem

ostatniego, który jest równy $n/(n+1)$, będą więc na pewno mniejsze od $n/(n+1)$. Wobec tego o sumach możemy powiedzieć, że dla $n \geqslant 7$ zachodzi nierówność (3), co należało udowodnić.

LEMAT 2. *Jeśli dla wyrazów jakiegoś ciągu nieskończonego o wyrazach dodatnich zachodzą nierówności*

$$b_2 < \lambda_1 b_1,\ b_3 < \lambda_2 b_2,\ \ldots,\ b_n < \lambda_{n-1} b_{n-1} \tag{7}$$

i jeśli iloczyn $\lambda_1 \ldots \lambda_n$ *przy* n *zmierzającym do* ∞ *dąży do* 0, *to ciąg* $\{b_n\}$ *dąży do* 0.

Dowód. Na podstawie nierówności (7) możemy napisać

$$b_2 < \lambda_1 b_1,\ b_3 < \lambda_1 \lambda_2 b_1,\ \ldots,\ b_n < \lambda_1 \lambda_2 \ldots \lambda_{n-1} b_1.$$

Ciąg po prawej stronie przy $n \to \infty$ zmierza do zera, bo b_1 jest wielkością skończoną, a iloczyn $\lambda_1 \ldots \lambda_n$ z założenia dąży do zera. Wyrazy ciągu b_n spełniają nierówności $0 < b_n < \lambda_1 \lambda_2 \ldots \lambda_{n-1} b_1$, a więc na podstawie twierdzenia o trzech ciągach ([5], str. 29) możemy powiedzieć, że ciąg b_n zmierza do zera, co mieliśmy pokazać.

Ciąg (2) spełnia założenia lematu 2, a wobec tego też dąży do zera. Dla $n \geqslant 7$ jako λ_n możemy przyjąć $n/(n+1)$. Obliczmy wartości λ_i dla $i < 7$:

$$\frac{a_3}{a_2} = \frac{5}{3},\ \frac{a_4}{a_3} = 1,\ \frac{a_5}{a_4} = \frac{21}{25},\ \frac{a_6}{a_5} = \frac{409}{504},\ \frac{a_7}{a_6} = \frac{2365}{2863}.$$

Możemy przyjąć, że $\lambda_4 = \lambda_5 = \lambda_6 = 1$; nierówności (7) będą zatem prawdziwe. Zamiast λ_2 musimy przyjąć liczbę większą od 5/3, a λ_3 ma być większe od 1. Przyjmijmy dla prostoty, że $\lambda_2 = 2$, $\lambda_3 = 8/7$ i wypiszmy iloczyn

$$\lambda_2 \ldots \lambda_n = 2 \cdot \frac{8}{7} \cdot 1 \cdot 1 \cdot 1 \cdot \frac{7}{8} \cdot \frac{8}{9} \cdot \ldots \cdot \frac{n}{n+1} = \frac{2 \cdot 8}{n+1} = \frac{16}{n+1};$$

przy $n \to \infty$ dąży on do zera.

Stosunek ilości różnych cykli do ilości wszystkich permutacji z n elementów, przy n zmierzającym do ∞, dąży do zera, ale bardzo wolno. Oznacza to, że dla małych n oszczędność rachunków jest niewielka. Dla dokładnego uporządkowania tablicy musimy jednak zbadać wszystkie cykle.

Do naszych celów nie jest potrzebne aż tak dokładne uporządkowanie i proponujemy wyznaczyć optymalny przydział znaków literom metodą przybliżoną, podaną niżej.

Uporządkujmy wiersze macierzy T_1 według malejących sum ich wyrazów, a kolumny macierzy T_2 według rosnących sum ich wyrazów. Dla jednoznaczności proponujemy w wypadku równych sum dwóch kolumn (wierszy) porównać wyrazy stojące w tym samym wierszu (ko-

lumnie) i przyjąć dla sum znak nierówności, jaki będzie prawdziwy dla pierwszej pary różnych wyrazów. Po uporządkowaniu wymnóżmy te macierze przez siebie i macierz iloczynu oznaczmy przez T_3^*. Szukamy takiego uporządkowania kolumn tablicy T_3^*, żeby suma wyrazów głównej przekątnej była minimalna. Jeśli zamienimy kolumnę o numerze k z kolumną o numerze l, to efekt tej zamiany obliczymy z wzoru

$$Z_{kl} = (a_{kl} - a_{kk}) + (a_{lk} - a_{ll}), \tag{9}$$

gdzie a_{ij} jest elementem tablicy T_3^*.

Jeśli wyrażenie (9) będzie większe od zera dla wszystkich par (k, l), to takiego uporządkowania nie polepszymy zamianą dwóch kolumn. Takie uporządkowanie będziemy uważali za optymalne. Jeśli Z_{kl} będzie większe od zera dla pewnych k i l, to możemy polepszyć uporządkowanie przez zamianę kolumny o numerze k z kolumną o numerze l, co też należy wykonać.

Uporządkowanie optymalne łatwo znajdziemy za pomocą tablicy T_4, którą otrzymamy z tyblicy T_3^*, odejmując od każdego elementu jej wiersza wyraz stojący na głównej przekątnej w danym wierszu. Jeśli przez a_{ij} oznaczymy elementy tablicy T_3^*, a przez b_{ij} elementy tablicy T_4, to $b_{ij} = a_{ij} - a_{ii}$, a wówczas

$$Z_{kl} = b_{kl} + b_{lk}.$$

Jeśli suma wyrazów symetrycznych w tablicy T_4 będzie dodatnia, to dane uporządkowanie jest już dobre. Dla jednoznaczności proponujemy najpierw wymieniać pary kolumn dające największy efekt.

Na materiale z punktu 3 spróbowaliśmy wyznaczyć dla języka polskiego alfabet o największej efektywności, korzystając z tworzywa stenograficznego JSSP.

W niektórych systemach stenograficznych niektóre znaki zmieniają sens wraz ze zmianą wielkości, co stwarza możliwość nieporozumień przy odczytywaniu stenogramów. Jeśli weźmiemy zbyt dużo znaków, dla każdej litery oddzielny znak, to siłą faktu do naszego tworzywa wejdą znaki bardziej złożone. Uważamy wobec tego, że warto sprawę tak uprościć, żeby podobnym literom (fonetycznie podobnym) przydzielić znaki różniące się tylko wielkością — podobnie jak to jest w stenografii Gregga. Gdy połączymy spółgłoski fonetycznie podobne, to okaże się, że różnych spółgłosek w języku polskim jest tylko 11. Jak już zaznaczyliśmy, w niektórych systemach stenograficznych [6] samogłoski są symbolizowane, wobec tego musimy mieć tylko 11 różnych znaków stenograficznych. Jako najprostsze i najbardziej różniące się między sobą wybraliśmy znaki podane w tablicy 5 i policzyliśmy szybkość pisania tych znaków. Wiadomo, że najszybciej pisze się kreskę prawoskośną. Przyjmujemy czas napi-

sania tej kreski za jednostkę, nazywając ją tempem (1 tempo = 1/160 min.). Za pomocą prób wyznaczono czas pisania innych znaków stenograficznych, z uwzględnieniem topografii danego znaku. Wyniki podajemy w tablicy 5. W tablicy 6 przedstawiamy statystykę topograficzną liter

TABLICA 5. Czas pisania znaków stenograficznych wyrażony w tempach

Znak	Na początku wyrazu	W środku wyrazu	Na końcu wyrazu
	1,2	1,2	1,2
	1,2	1,2	1,2
	1,0	1,4	1,4
	1,4	1,4	1,4
	1,4	1,4	1,4
	1,4	1,4	1,4
	1,4	1,4	1,4
	1,5	1,6	1,6
	1,6	1,7	1,6
	1,6	1,7	1,6
	1,8	2,0	2,0

opartą na materiale z punktu 3. Z materiału tego usunięto częste wyrazy, rdzenie i przedrostki, a mianowicie: bez (4 razy wystąpiło we wszystkich tekstach), jednak (6 razy), pod (7 razy), przed (7 razy), dal (11 razy), jego (11 razy), jak (16 razy), jest (18 razy), któr (33 razy). Połączono litery fonetycznie podobne w grupy. Np. litery *S*, *Ś*, *Z*, *Ź*, *Ż* traktowano jako jedną literę. Ponieważ *Ż* nie różni się fonetycznie od *RZ*, od częstości litery *R* odjęto częstość grupy *RZ*. Ponieważ połączono *H* i *CH* w jedną grupę, od częstości litery *C* odjęto częstość grupy *CH*. *C* połączono

TABLICA 6. Frekwencje liter w czterech tekstach razem z uwzględnieniem topografii

Litera	Osobno	Na początku wyrazu	W środku wyrazu	Na końcu wyrazu	Razem
A	19	56	1590	494	2159
Ą	—	—	184	91	275
B	—	103	178	5	286
C	—	114	800	44	958
Ć	—	1	3	66	70
D	—	209	594	60	863
E	—	20	1261	488	1769
Ę	—	—	169	128	297
F	—	38	29	—	67
G	—	71	248	11	330
H	—	10	79	157	246
I	115	35	1535	261	1946
J	—	115	237	126	478
K	—	173	661	76	910
L	—	50	408	8	466
Ł	—	10	378	154	542
M	—	150	302	179	631
N	—	285	1050	21	1356
Ń	—	—	28	26	54
O	27	218	1263	338	1846
Ó	—	4	223	—	227
P	—	480	252	4	736
R	—	133	932	28	1093
S	—	315	710	13	1038
Ś	—	30	118	9	157
T	—	202	620	58	880
U	—	117	338	229	684
W	155	329	668	74	1226
Y	—	—	708	350	1058
Z	60	306	1016	93	1475
Ż	—	49	107	32	188
Ź	—	1	28	1	30
Razem	376	3624	16717	3624	24341

z *CZ* więc od częstości litery *Z* odjęto częstość grupy *CZ*. Pisanie znaku oddzielnie jako samodzielne słowo zajmuje tyle czasu, co napisanie tego znaku na początku wyrazu, wobec tego w tablicy 6 dodano do siebie dwie pierwsze kolumny. Wyniki przedstawiono w tablicy 7.

Jako prawdopodobieństwo pojawienia się danej litery w tekście przyjęliśmy frekwencję tej litery z tablicy 7 podzieloną przez ilość liter podanych w tej tablicy. Jako czas napisania danego znaku stenograficznego wzięliśmy dane z tablicy 5. Metodą przybliżoną opisaną wyżej przy-

TABLICA 7. Frekwencje liter z grup liter fonetycznie podobnych.

Grupa liter	Na początku lub osobno	W środku wyrazu	Na końcu wyrazu	Razem
S,Ś,Z,Ź,Ż,RZ	733	1775	136	2644
T,D	395	1168	93	1656
N,Ń	285	1072	47	1404
W,F	522	697	74	1293
K,G	211	898	65	1174
P,B	565	430	9	1004
L,Ł	60	770	162	992
C,Ć,CZ	85	597	110	792
R	126	627	23	776
M	150	302	179	631
CH,H	10	79	157	246
Razem	3142	8415	1055	12612

TABLICA 8

Litera	Znak				
					
S,Ś,Z,Ź,Ż,RZ	0	0	235,6	528,8	528,8
D,T	0	0	173,2	331,2	331,2
B,P	25,2	25,2	0	226,0	226,0
F,W	−258,6	−258,6	−208,8	0	0
N,Ń	−280,8	−280,8	−114,0	0	0
G,K	−234,8	−234,8	−84,4	0	0
L,Ł	−198,4	−198,4	−24,0	0	0
R	−297,8	−297,8	−193,0	−142,6	−142,6
C,Ć,CZ	−376,5	−376,5	−252,1	−218,1	−218,1
M	−282,6	−282,6	−216,4	−156,4	−156,4
CH,H	−194,8	−194,8	−149,6	−145,6	−145,6

TABLICA 9

Litera	Znak				
S,Ś,Z,Ź,Ż,RZ	3172,8	3172,8	3408,4	3701,6	3701,6
D,T	1987,2	1987,2	2160,4	2318,4	2318,4
B,P	1204,8	1204,8	1179,6	1405,6	1405,6
F,W	1551,6	1551,6	1601,4	1810,2	1810,2
N,Ń	1684,8	1684,8	1851,6	1965,6	1965,6
G,K	1408,8	1408,8	1559,2	1643,6	1643,6
L,Ł	1190,4	1190,4	1364,8	1388,8	1388,8
R	931,2	931,2	1036,0	1086,4	1086,4
C,Ć,CZ	950,4	950,4	1074,8	1108,8	1108,8
M	757,2	757,2	823,4	883,4	883,4
CH,H	295,2	295,2	340,4	344,4	344,4

dzieliliśmy znaki stenograficzne spółgłoskom, zgodnie z porządkiem kolumn w tablicy 8. Wyniki podano w tablicy 9 (zamiast prawdopodobieństw brano częstości liter, co zostało uwzględnione w końcowych obliczeniach). Przyjmując czas pisania równy 0 dla liter *I* i *Y* pisanych w środku wyrazu, równy 1 tempo dla samogłosek *A*, *E*, *O*, *Ó*, *U* oraz dla *I* i *Y* pisanych na początku i na końcu wyrazu, a 1, 3 tempa dla *Ą*, i *Ę*, otrzymamy dla naszego alfabetu

$$S = 1{,}225 \quad \text{i} \quad E = 0{,}816.$$

Te same wyrażenia policzone dla alfabetu JSSP w przybliżeniu są równe

$$S = 1{,}426 \quad \text{i} \quad E = 0{,}701.$$

A oto alfabet stenograficzny, jaki można zaproponować, opierając się na wynikach z tablicy 9:

TABLICA 8 c.d.

Znak					
					
528,8	528,8	984,3	1235,1	1235,1	1968,6
331,2	331,2	622,9	779,2	779,2	1245,8
226,0	226,0	370,3	469,8	469,8	715,4
0	0	206,4	328,3	328,3	671,4
0	0	252,3	388,0	388,0	785,4
0	0	213,7	324,6	324,6	662,2
0	0	192,4	275,4	275,4	583,2
−142,6	−142,6	0	75,3	75,3	297,8
−218,1	−218,1	−68,1	0	0	240,1
−156,4	−156,4	−45,2	0	0	192,2
−145,6	−145,6	−97,4	−88,5	−88,5	0

TABLICA 9. c.d.

Znak					
					
3701,6	3701,6	4157,1	4407,9	4407,9	5141,4
2318,4	2318,4	2610,1	2766,4	2766,4	3233,0
1405,6	1405,6	1549,9	1649,4	1649,4	1895,0
1810,2	1810,2	2016,6	2138,5	2138,5	2481,6
1965,6	1965,6	2217,9	2353,6	2353,6	2751,0
1643,6	1643,6	1857,3	1968,2	1968,2	2305,8
1388,8	1388,8	1581,2	1664,2	1664,2	1972,0
1086,4	1086,4	1229,0	1304,3	1304,3	1526,8
1108,8	1108,8	1258,7	1326,9	1326,9	1567,0
883,4	883,4	994,6	1039,8	1039,8	1232,0
344,4	344,4	392,6	401,5	401,5	490,0

Proponowany alfabet stenograficzny

Nr	Litera	Znak stenograficzny	Nr	Litera	Znak stenograficzny
		Spółgłoski i częste(1) grupy spółgłoskowe			
1	*B*		16	*L*	
2	*P*		17	*Ł*	
3	*C = Ć*		18	*M*	
4	*CZ*		19	*R*	
5	*CH = H*		20	*PR*	
6	*D*		21	*PRZ*	
7	*T*		22	*S = Ś*	
8	*DZ*		23	*SZ*	
9	*DŹ*		24	*Z = Ź*	
10	*W*		25	*Ż = RZ*	
11	*F*		26	*SK*	
12	*N = Ń*		27	*SC = ŚĆ*	
13	*WN*		28	*CZN*	
14	*K*		29	*J*	
15	*G*				

(1) Grupy spółgłoskowe o częstości większej niż 0,5 %.

Proponowany alfabet stenograficzny c.d.

Nr	Litera	Znak stenograficzny	Nr	Litera	Znak stenograficzny
		Samogłoski (pisane w górę lub w prawo)			
30	*A*	/	34	*Ą*	~
31	*E*	/	35	*Ę*	~
32	*O*	—	36	*I* = *Y*	/ na końcu
33	*Ó* = *U*	—			/ na początku

Znak *i-y* piszemy tylko na początku i na końcu wyrazu. Litera *J* została potraktowana specjalnie, ponieważ fonetycznie jest podobna do *I*, więc jest to coś pośredniego między samogłoską i spółgłoską.

Znaczniki dla częstych wyrazów ([2]), częstych rdzeni ([3]) i przedrostków

Nr	Wyraz	Znacznik	Nr	Wyraz	Znacznik
1	jest	/	6	jednak	~
2	jak	ρ	7	dla	δ
3	jego	ρ	8	bez	~
4	pod	ʔ	9	dal	δ
5	przed	ℓ	10	któr	—

Zostało jeszcze do rozstrzygnięcia, czy znaki stenograficzne zostały wybrane najlepiej. Wyznaczenie optymalnego tworzywa stenograficznego wymaga specjalnych badań, a tego nikt dotychczas nie robił. Orzeczenie, który znak jest bieglejszy, jest subiektywne. Jedni uważają np.,

([2]) Wyrazy, które wystąpiły w każdym tekście przynajmniej raz.

([3]) Rdzenie i przedrostki, które po usunięciu częstych wyrazów wystąpiły w każdym tekście przynajmniej raz.

że bieglejsze są koła lub pętle pisane w prawą stronę, inni — że bieglejsze są te znaki pisane w lewą stronę. Jak jest naprawdę, można orzec dopiero po przeprowadzeniu wielu prób na różnych osobach.

7. Uwagi końcowe. Szybkość stenografowania chyba w małym stopniu zależy od dobrego alfabetu, nikt bowiem nie stenografuje pisząc literę za literą. Największe uproszczenie dają skróty wyrazów i zwrotów. Jak wybrać częste wyrazy, dla których warto wprowadzić specjalne oznaczenia? B. Epstein w pracy [3] mówi o badaniach częstości słów ważnych (to znaczy słów, których sens powinien umieć rozpoznać student po dwu latach uczenia się danego języka) przeprowadzonych dla języka portugalskiego w Brazylii [1]. Badanie to wykazało, że „wiele ważnych słów dla języka portugalskiego w Brazylii występuje tylko raz lub dwa razy na 100000 słów. Jest to niewątpliwie prawdą także dla innych języków". W pracy tej jest też podane rozumowanie prowadzące do ustalenia liczności próbki, która pozwoliłaby odróżnić słowo B_2 występujące 2 razy w 100000 od słowa B_1 występującego raz w 100000. Wyliczono, że do tego celu należy użyć próbki liczącej z grubsza 1000000 wyrazów. Wydaje się nam, że dla wyłowienia częstych wyrazów w języku polskim należy wziąć kilka próbek po 100000 wyrazów z tekstów o różnej tematyce i za częste uznać te wyrazy, które występują we wszystkich tekstach przynajmniej raz, za częste rdzenie, przedrostki i przyrostki — te, które po usunięciu częstych wyrazów, wystąpią we wszystkich tekstach przynajmniej raz. Dla próby wybraliśmy częste wyrazy, rdzenie i przedrostki z naszego materiału. Oto one: bez, dal, dla, jak, jednak, jego, jest, któr, pod i przed. Można się spodziewać, że częstych rzeczowników i czasowników w języku polskim jest mało, natomiast w każdym tekście z osobna są częste rzeczowniki, czasowniki i inne części mowy, ściśle związane z tematem. W tekście handlowym często powtarza się słowo „zysk", które w pozostałych tekstach nie wystąpiło wcale. Najczęstsze, wspólne dla wszystkich tekstów, słowo „jest" wystąpiło 18 razy (na 4000 wyrazów), a słowo „zysk" w tekście handlowym wystąpiło 34 razy (na 1000 wyrazów). Które z tych słów jest częstsze w języku polskim? Na pewno częstsze jest słowo „jest".

Jakie płyną stąd wnioski? Oczywisty wydaje się wniosek, że nie może być jeden system stenograficzny. Może być jeden alfabet wspólny dla wszystkich specjalności i niezbyt obszerny system skrótów, ale musi być opracowany specjalnie dla danej tematyki system znaczników dla wyrazów tej czy innej specjalności. A więc stenografia handlowa, administracyjna, parlamentarna itd.

Oczywiście, dla przydzielenia specjalnych znaków wyrazom nie jest potrzebne uporządkowanie wyrazów według częstości. Ważne jest raczej, żeby przydzielony znacznik dał się łatwo zapamiętać, musi więc

być on podobny do niektórych znaków danego wyrazu napisanego alfabetem stenograficznym czy łacińskim. Znak, którego nie potrafimy powiązać z wyrazem, trudno będzie zapamiętać. A w ogóle, jak dużo znaków można zapamiętać? Współczesne pismo chińskie używane potocznie zawiera 3000-6000 znaków. Dla stenografowania potocznego języka polskiego trzeba, zdaniem fachowców, opanować co najmniej 1000 skrótów. Można się spodziewać, że rozsądna ilość skrótów, jaką musi rozporządzać dobry stenograf, jest zawarta między 1000 a 2000. Nie warto obciążać pamięci skrótami wyrazów występujących rzadko, dla częstych zaś wyrazów nie powinno zabraknąć prostych oznaczeń. I dlatego do tej ilości znaków, na jaką się zdecydujemy, należy umiejętnie wybrać wyrazy.

Prace cytowane

[1] C. B. Brown, W. M. Carr i M. L. Shane, *A graded work book of Brasilian-Portuguese*, Crofts Co., New York 1945.

[2] H. Cramer, *Metody matematyczne w statystyce*, PWN, Warszawa 1958.

[3] B. Epstein, *Statistical aspects of the Russian word count*, rozdz. V, str. 23-25, w książce H. H. Josselson, *The Russian word count and frequency analysis of grammatical categories of standard literary Russian*, Wayne University Press, Detroit 1953.

[4] J. Kalkowski, *Stenografia*, Przekrój nr 1019 z dnia 18. X. 1964, str. 4.

[5] F. Leja, *Rachunek różniczkowy i całkowy*, PWN, Warszawa 1954.

[6] R. Łazarski i J. Kaczmarek, *Jednolity system stenografii polskiej*, część I, Warszawa 1958.

[7] A. Mostowski i M. Stark, *Elementy algebry wyższej*, PWN, Warszawa 1963.

[8] A. A. Schlichting, *A survey of shorthand systems*, Seattle 1949, niepublikowany maszynopis.

[9] M. Szejnert, *Co dalej skoropisie?* Polityka nr 53 (513) z dnia 31. XII. 1966, str. 10.

[10] K. Szuter, *Statystyka częstotliwości składników języka polskiego*, Materiały z zakresu teorii i praktyki stenotypii, zeszyt 1, Warszawa 1957, str. 4-34.

[11] C. A. Thomas, *Introductory readings in Thomas shorthand*, Prentice-Hall, New York 1938.

DZIAŁ ZASTOSOWAŃ PRZYRODNICZYCH, GOSPODARCZYCH I TECHNICZNYCH
INSTYTUTU MATEMATYCZNEGO PAN, WROCŁAW

Praca wpłynęła 12. 5. 1967,
nowa wersja 10. 5. 1969

Л. ЗУБЖИЦКА (Вроцлав)

ОБ ОПРЕДЕЛЕНИЮ СТЕНОГРАФИЧЕСКИХ СИСТЕМ

РЕЗЮМЕ

Работа касается задачи приспособления стенографической системы к языку, т.е. предложения такой системы, которая делалабы возможным самое быстрое писание текстов на данном языке. Такая система должна связывать скорость писания знаков с частотой их появления. В этой проблеме можно выделить две задачи: 1) Как избрать стенографические знаки в систему, 2) как поставить эти знаки в соответствие буквам, группам согласных, выражениям и т.п.

В работе рассмотрена только вторая задача. В связи с этим обсуждаются методы статистической выборки с целью определения частоты элементов языка. В п. 6 представлен метод определения оптимального стенографического алфавита для данного языка и данных стенографических знаков. Показано тоже новое соответствие между буквами и стенографическими знаками системы польской стенографии.

L. ZUBRZYCKA (Wroclaw)

ON THE DETERMINATION OF STENOGRAPHIC SYSTEMS

SUMMARY

The paper deals with the problem of adapting a stenographic system to the language, i.e. of providing such a system which would allow most rapid writing of texts in the given language. Such a system should link the speed of sign writing with the frequency of their occurrance. Two goals may be stated here: 1) the selection of stenographic signs to form a system, 2) the allocation of these signs to letters, consonant groups, words, etc.

The paper is devoted to the second goal. Sampling experiments for establishing the frequency of language elements are considered. In section 6 a method of determining the stenographic alphabet being optimal for a given language and a fixed set of stenographic signs is given. It is also shown how the elements of the Polish stenographic system should be assigned to letters.