3.10.13. — akustyka mowy

Krzysztof Kowalski

MATEMATYCZNE MODELOWANIE ROZKŁADU CZĘŚCI WYSTĘPOWANIA FONEMÓW JĘZYKA POLSKIEGO

15/1995

WARSZAWA 1995

ISSN 0208-5658

Praca wpłynęła do Redakcji dnia 13 lutego 1995 r.

Na prawach rękopisu

Instytut Podstawowych Problemów Techniki PAN
Nakład 100 egz. Ark. wyd. 0,75 Ark. druk. 1,0
Oddano do drukarni w kwietniu 1995 r.

Wydawnictwo Spółdzielcze sp. z o.o.
Warszawa, ul. Jasna 1

Krzysztof Kowalski
Zakład Fonetyki Akustycznej PAN
Poznań

MATEMATYCZNE MODELOWANIE ROZKŁADU CZĘSTOŚCI WYSTĘPOWANIA FONEMÓW JĘZYKA POLSKIEGO

Streszczenie

W pracy przybliżano matematycznie rozkład częstości występowania fonemów w języku polskim. Użyto następujących modeli matematycznych: rozkładu Zipfa, rozkładu liniowego, rozkładu logarytmicznego, wielomianów potęgowych, wielomianów Hermite'a-Czebyszewa oraz regresji z opóźnieniem. Najlepszy sposób przybliżenia uzyskano przy pomocy ostatniego modelu bo aż prawie w 100%. Omówiono konsekwencje wynikające z tych matematycznych modeli. Dotyczą one niezależności oraz zależności pojawiania się fonemów w języku polskim, a także takich charakterystyk jak redundacja oraz oszacownia największego słownika dla języka polskiego. Poczyniono również pewne spostrzeżenia dotyczące minimalnego słownika języka polskiego.

1.Wstęp

W pracy dokonano aproksymacji matematycznej częstościowego występowania fonemów języka polskiego przy pomocy różnych modeli matematycznych (rozkładu Zipfa, rozkładu liniowego, rozkładu logarytmicznego, wielomianów potęgowych, wielomianów Czebyszewa-Hermite'a). Najlepszy sposób przybliżenia otrzymano jednak w wyniku zastosowania tzw. regresji z opóźnieniem. Cechą charakterystyczną jej jest bardzo wysoki stopień "wytłumaczenia " (99.6 %) przy niewielkiej ilości wyrazów. Na podstawie obecnych badań można powiedzieć, że jest to cecha charakterystyczna dla języka polskiego niespotykana w innych językach (tj. angielskim, rosyjskim oraz

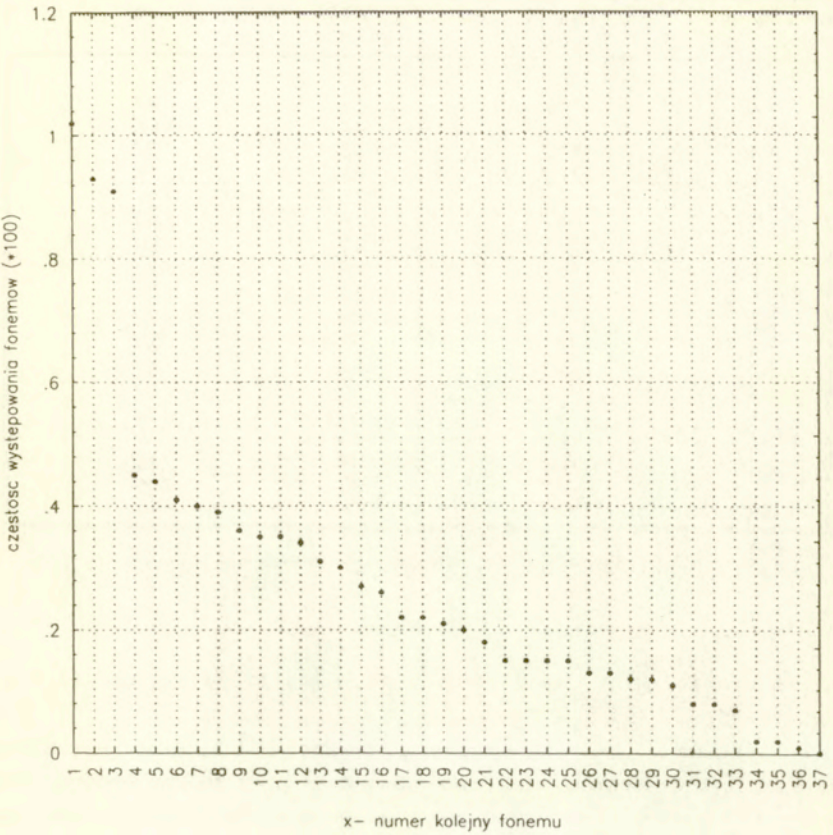
niemieckim (dla języka pisanego)) w takim stopniu. Wyniki te wskazują na sposób "związania" głosek w języku polskim (po około 3 fonemów).

2. Rozkład częstości występowania fonemów w języku polskim.

Jako materiał do modelowania matematycznego przyjęto rozkład częstości fonemów w języku polskim zamieszczony w książce Jassema [:1 str. 283]. Przedstawiono ten rozkład w sposób graficzny na rys. 1. Do obliczeń przyjęto rozkład w skali takiej jak na rys. 1. tzn. f(x):100 (ze względu na uniknięcie ewentualnych trudności obliczeń). Przedstawiony rozkład został sporządzony z dokładnością 1 promilla. Liczba 0 przy fonemie x=37 oznacza, że jego częstość względna jest mniejsza niż 0.5 promilla. Uzyskanie takiej dokładności wymagało opracowania tekstów, w których łączna ilość fonemów wynosiła 50000. Znaczenie x- numerów kolejnych fonemów polskich wyjaśniono w tabeli 1. (np. dla x=23 odpowiada fonem /ts/- patrz strzałka w tabeli).

Przedstawiony rozkład częstości fonemów polskich jest rozkładem dyskretnym określonym na skończonym przedziale. Modelowanie matematyczne pozwala dokonać dwóch operacji abstrakcyjnych a mianowicie: przedstawić rozkład w formie ciągłej oraz określić go w nieskończonym przedziale. Możliwość dokonania wspomnianych wyżej pozwala przynajmniej teoretycznie rozważać zagadnienia dalszego podziału fonemów (np. na allofony) oraz zagadnienia nieskończonego alfabetu. Te ostatnie zagadnienia nie są przedmiotem tej pracy, ale warto o nich pamiętać.

Cechą charakterystyczną rozkładu występowania fonemów polskich jest nagłe "wyniesienie" trzech najczęściej występujących fonemów, a mianowicie: /e/, /a/, /o/. Jest to cecha charakterystyczna dla języka polskiego. Tą cechą przypomina też język "pisany" angielski w "wyniesieniu" znaku /e/. Takiej cechy na przykład nie posiada język "pisany" rosyjski.



Rys.1. Rozkład częstości występowania fonemów w języku polskim według Jassema [1: str. 283].

3. Modelowanie matematyczne rozkładu częstości występowania fonemów w języku polskim za pomocą regresji z opóźnieniem.

Analiza regresji z opóźnieniem jest narzędziem dla badania zależności występujących między zmiennymi będącymi w opóźnieniu do siebie. Takie metody dotychczas stosowano w ekonometrii np. dla badania związków między inwestycjami w maszyny a dochodami po

Numer kolejny	Fonem	Numer kolejny	Fonem
1	/e/	20	/ʃ/
2	/a/	21	/z/
3	/o/	22	/ɕ/
4	/j/	23	/ʦ/
5	/t/	24	/f/
6	/ɨ/	25	/g/
7	/n/	26	/b/
8	/i/	27	/ʨ/
9	/r/	28	/ʒ/
10	/m/	29	/ʧ/
11	/v/	30	/x/
12	/u/	31	/ʤ/
13	/p/	32	/ŋ/
14	/s/	33	/c/
15	/k/	34	/ʑ/
16	/ɲ/	35	/ʣ/
17	/d/	36	/ɬ/
18	/w/	37	/ʤ/
19	/l/		

Tabela 1. Znaczenie numerów kolejnych x występujących na rys. 1.

jakimś czasie, który na podstawie tych analiz da się określić [4: str. 683-690], [7].

Przypuśćmy, że mamy zależną zmienną Y(t) oraz niezależną zmienną L(t). Zmienna zależna nazywana jest także endogeniczną (ang. endogenous). Zmienna niezależna określana jest także jako exogeniczna (ang. exogenous) [4: str. 683].

Sposobem opisania tego typu związków jest podanie prostego liniowego związku pomiędzy zmiennymi Y(t) oraz L(t), który dalej określimy.

Dwa procesy Y(t) oraz L(t) przybliża się z następujący sposób:

$$Y(t) = \sum_{i=0}^{i=p} c_i * L(t-i) \qquad (1)$$

gdzie: c_i są współczynnikami określonymi metodą najmniejszych kwadratów ze współczynnikiem determinacji R^2 [2: str. 265, 267]:

$$R^2=\frac{(n*\sum_{i=0}^{i=n} f(x_i)*g(x_i)-\sum_{i=0}^{i=n} f(x_i)*\sum_{i=0}^{i=n} g(x_i))^2}{(n*\sum_{i=0}^{i=n} f(x_i)^2-(\sum_{i=0}^{i=n} f(x_i))^2)*(n*\sum_{i=0}^{i=n} g(x_i)^2-(\sum_{i=0}^{i=n} g(x_i))^2)}$$

gdzie:

$g(x_i)$- funkcja wynikająca z modelowania matematycznego

$f(x_i)$- funkcja rozkładu występowania fonemów polskich

n - liczba par $\{g(x_i), f(x_i)\}$

p- jest współczynnikiem

W przypadku wzoru (1) należy Y(t) utożsamić z rozkładem częstości występowania fonemów polskich f(x) natomiast L(t) z numerem kolejnym fonemu według częstości ich występownia oznaczonym jako x.

Wyniki przybliżenia funkcji rozkładu częstości występowania fonemów w języku polskim według wzoru (1) są następujące:

c_0=0.135339

c_1=0.04290

c_2=-0.049531

c_3=-0.141967

Współczynnik determinacji wynosi 99.6 %. Podane współczynniki są istotne statystycznie. Parametr p wynosi 3. Wyższe wartości p nie poprawiają już współczynnika determinacji.

Obliczenia przeprowadzono wykorzystując program CSS: Statistica (procedury "distributed lags") [4].

4. Inne modele matematyczne odwzorowujące rozkład częstości występowania fonemów w języku polskim.

W pracy zbadano także metodą najmniejszych kwadratów [2: str. 96] inne modele matematyczne mogące również być odzwierciedleniem rozkładu częstości występowania fonemów polskich takich, jak:

- rozkład Zipfa:

$$f(x) = k_1 * x^p, \tag{2}$$

gdzie:

$k_1 = 1.206$

$p = -0.603$

Współczynnik determinacji wynosi 88%.

-przybliżanie funkcją liniową:

$$f(x) = r_0 + r_1 * x, \tag{3}$$

gdzie:

$r_0 = 0.63715$

$r_1 = -.0193$

Podane współczynniki są istotne statystycznie. Współczynnik determinacji wynosi około 74%.

-przybliżanie funkcją logarytmiczną:

$$f(x) = d_0 + d_1 * \log(x), \tag{4}$$

gdzie:

$d_0 = 0.996892$

$d_1 = -.270551$

Podane współczynniki są istotne statystycznie. Współczynnik determinacji wynosi około 94%.

- przybliżanie funkcjami potęgowymi:

$$f(x) = \sum_{i=0}^{n1} b_i * x^i, \tag{5}$$

gdzie:

$b_0 = 1.318074$

$b_1 = -0.27369$

$b_2 = .029552$

$b_3 = -0.001561$

$b_4 = .000039$

$b_5 = -0.$

Najwyższy stopień aproksymowanego wielomianu wynosi n1=5. Przedstawione współczynniki są istotne statystycznie. Współczynnik determinacji wynosi 95%.

- przybliżanie funkcjami Czebyszewa-Hermite'a:

$$f(x) = \sum_{i=0}^{i=n2} a_i * H_i \qquad (6)$$

$$H_0(x) = 1$$

$$H_2(x) = x$$

$$H_2(x) = x^2 - 1$$

$$H_3(x) = x^3 - 3*x$$

$$H_4(x) = x^4 - 6*x + 3,$$

gdzie: a_0=1.184015
a_1=-.177246
a_2=.012558
a_3=-.000391
a_4=.000004

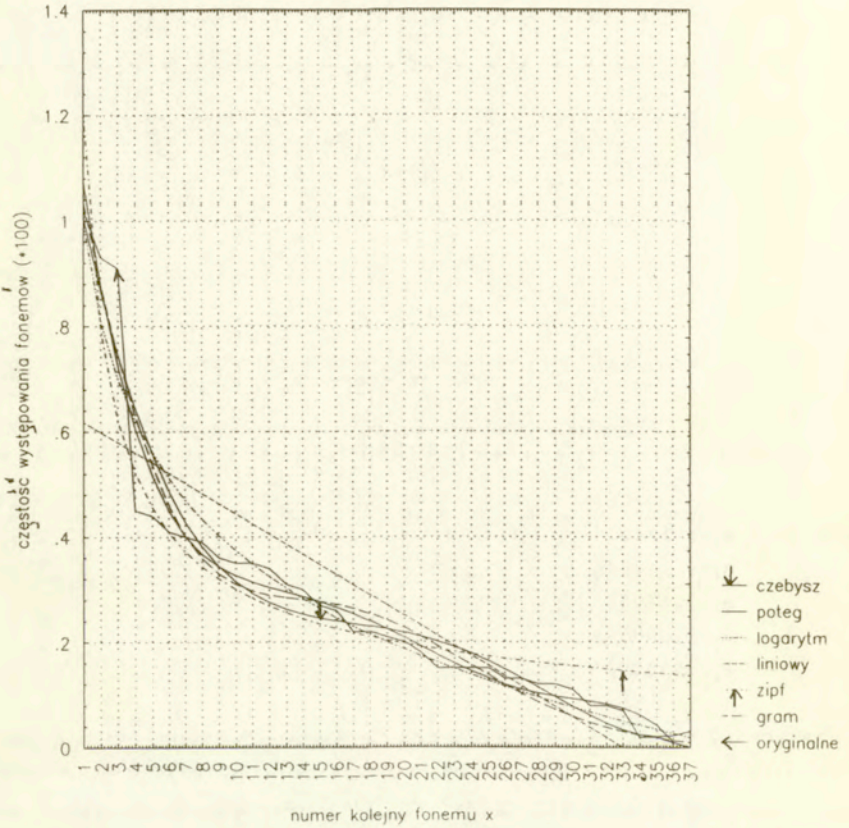
Parametr n2 wynosi 4. Współczynnik determinacji wynosi około 95%.

-przybliżanie poprzez szereg Grama-Charliera typu A [2: str. 214]:

$$f(x) = f_0 * \exp(f_1 * x^2 + f_2) * (1 - f_3 * H_1 + f_4 * H_2 + f_5 * H_3 + f_6 * H_4), \qquad (7)$$

gdzie: f_0=0.421686
f_1=-0.003897
f_2=1.062003
f_3=0.148936
f_4=0.010254
f_5=-0.000157
f_6=-0.000000

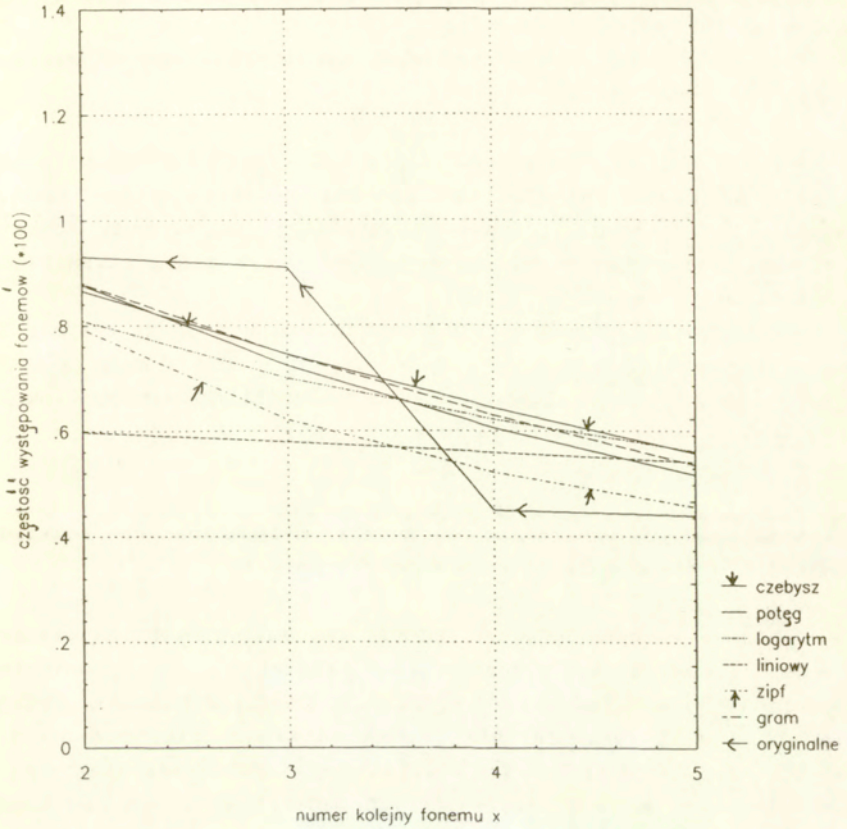
Poprzez H_i oznaczono wielomiany Czebyszewa-Hermite'a, które zdefiniowane zostały powyżej. Współczynnik determinacji wynosi 95%.



Rys. 2. Rysunek przedstawiający w sposób sumaryczny wyniki uzyskane w pracy dla różnych modeli matematycznych.

Na rysunkach 2 oraz 3 przedstawiono w sposób sumaryczny wyniki aproksymacji przy czym przyjęto następujące oznaczenia:

"oryginalne" -to dane oryginalne i model matematyczny regresji z opóźnieniem (z powodu bardzo małych różnic



Rys. 3. Rysunek przedstawiający "skok" pomiędzy 3 i 4 fonemem dla różnych modeli matematycznych.

utożsamiono oba wykresy, różnice wynoszą średnio około 0.002)

" zipf" -to model matematyczny oparty na rozkładzie Zipfa

"liniowy" -to model matematyczny oparty o funkcję liniową

"logarytm" -to model matematyczny oparty o funkcję logarytmiczną

"potęg" -to model matematyczny oparty o funkcje potęgowe

"czebysz" -to model matematyczny oparty na wielomianach Czebyszewa-Hermite'a.

"gram" -to model matematyczny oparty na szeregu Grama-Charliera typu A

Rozkład liniowy oraz rozkład Zipfa słabo aproksymują empiryczną funkcję rozkładu występowania fonemów polskich. Inne modele matematyczne w miarę jednakowo przybliżają empiryczną funkcję rozkladu. Współczynnik determinacji jest dobrą miarą określającą stopień przybliżenia.

Skok pomiędzy fonemami 3 i 4 przedstawiony jest na rysunku 3. Przybliżenia dla fonemu x=3 są poniżej oraz dla fonemu x=4 są powyżej danych oryginalnych. Różnice dochodzą do wartości około 0.32 .

5. Konsekwencje wynikające z modeli matematycznych rozkładu częstości występowania fonemów w języku polskim.

Spośród różnych użytych modeli matematycznych największy stopień aproksymacji wykazuje tzw. regresja z opóźnieniem przybliżająca rozkład w 99.6 %. Taki stopień przybliżenia wydaje się być charakterystyczny dla języka polskiego, niespotykany dla języków "pisanych" takich jak rosyjski, angielski oraz niemiecki w takim stopniu (na podstawie własnych badań). Na razie nie znane jest genetyczne wytłumaczenie tego wyniku. Można jednak przypuszczać, że częstości występowania fonemów w języku polskim nie występują niezależnie, lecz pojawiają się w pewnych "typowych" połączeniach,które tworzą najczęściej występujące fonemy jak: e, a, o. Taki wynik sugeruje też możliwość bardziej "ekonomicznego" kodowania opartego nie o fonemy, ale o kombinacje np. trzech fonemów. Być może sposób wyróżnienia tych kombinacji może być oparty na "rdzeniach" nazw, nazwisk, słów lub mógłby być połączony z "podobieństwem" fonetycznym. Modele matematyczne rozkładu częstości fonemów polskich pozwalają za pomocą skończonej liczby fonemów wnosić o zagadnieniach "nieskończonych". Niech takim przykładem będzie podanie teoretycznie największej liczby słów dla języka polskiego, która wynosi około 62 miliardy słów przyjmując charakterystyki statystyczne występowania fonemów w języku polskim,

przyjmując 7.5 fonema na słowo oraz przyjmując redundancję równą zeru. Wyraża to symbolicznie wzór (8):

$$H_{37} \approx H_{\infty} \Leftrightarrow H_{24.78 \cdot 7.5} \approx H_{\infty}, \tag{8}$$

gdzie:

H-oznacza entropię definiowaną tak jak w teorii informacji [1: str. 291].

Przy wyznaczeniu tej liczby korzystano z następującego rozumowania:

$$H_{fonemów} = \frac{H_{słów}}{k}, \tag{9}$$

gdzie:

k- to ilość fonemów przypadających na słowo.

Parametr k przyjęto jako :

$$k = \frac{H_{słów}}{H_{fonemów}} = 7.5 \tag{10}$$

Zakładając dalej, że:

$$H_{słów} * (1-Z) = \log_2(n) \tag{11}$$

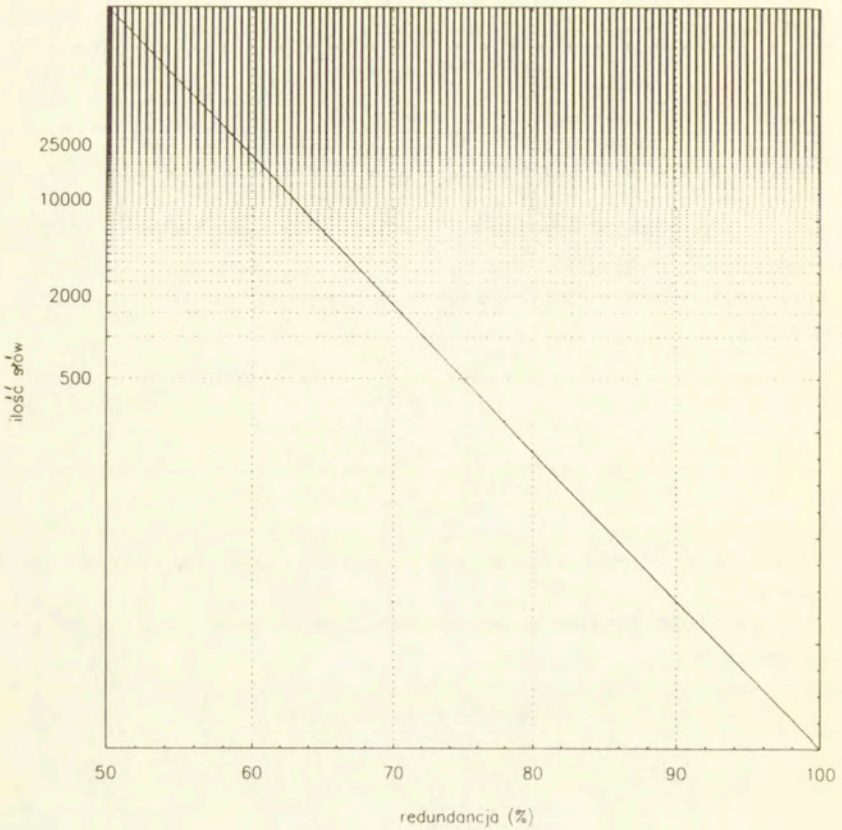
otrzymujemy:

$$n = 2^{H_{fonemów} * k * (1-Z)}, \tag{12}$$

gdzie:

Z-oznacza redundancję [1: str. 297], [12: str. 602].

Zależności (9) oraz (11) zaczerpnięto z pracy Głuszkowa [12: str. 602].



Rys. 4. Ilustracja ilości słów w zależności od redundancji

Z pewną ostrożnością należy traktować informacje o tym jakoby w myśl prawa Zipfa nieduża ilość słów (np. około 1000 słów) miała dobrze przybliżać całe słownictwo języka polskiego lub innego języka (np. [9: str. 191-192]). Redundancja w takim przypadku w języku polskim powinna wynosić około 72% (rys. 2). Oznacza to, że w języku polskim mówionym redundancja jest mniejsza np. 60% dla około 20000 słów lub 50% dla około 250000 słów. Oszacowanie maksymalnej ilości słów dla języka polskiego jest niezbędne dla oszacowań minimalnego słownika.

6.Wnioski

Dysponowanie modelem matematycznym rozkładu częstości występownia fonemów polskich pozwala poznać ich "strukturę matematyczną". Uwidacznia to także nieznane dotąd cechy charakterystyczne np. za pomocą regresji z opóźnieniem. Ten ostatni model matematyczny posiada tak dużą dokładność przybliżenia średniokwadratowego, że aproksymacja jest jednocześnie praktycznie interpolacją. Taki przypadek jest niezwykle rzadki w naukach eksperymentalnych. Jest to także sposób przedstawienia języka polskiego w postaci matematycznych niezmienników. Poznanie tego typu cech może być pomocne w zagadnieniach rozpoznawania mowy. Możliwy jest też teoretycznie sposób rozpoznawania mowy oparty na częstościach występowania fonemów podobnie jak to czynią kryptolodzy. Taka możliwość związana jest jednak z olbrzymią pracochłonnością oraz olbrzymim kosztem finansowym.

Wyniki te wskazują także to,że proces mowy jest z punktu widzenia matematycznego w znacznej części procesem autoregresji.

Bardzo ważną sprawą jest uczynić także spostrzeżenie, aby słowniki frekwencyjne dawały wyobrażenie o rzeczywistych własnościach językowych (w postaci charakterystyk matematycznych, takich np. jak redundancja), zamiast traktować prawo Zipfa jako rodzaj magicznego prawa mającego umotywować ich użyteczność [11: str. 7]. Autor nie neguje wartości słowników frekwencyjnych, ani prawa Zipfa, chodzi jedynie o ich właściwą interpretację.

Bibliografia

[1] W. Jassem, Podstawy fonetyki akustycznej, PWN , Warszawa, 1973.

[2] M.G. Kendall, W.R. Buckland, Słownik terminów statystycznych", Państwowe Wydawnictwo Ekonomiczne, Warszawa, 1986.

[3] I.N. Bronsztejn, K.A. Siemiendajew "Matematyka-Poradnik Encyklopedyczny, Warszawa, 1968.

[4] CSS:STATISTICA (Volume II): opis pakietu statystycznego CSS.

[5] P.A. Frost, Some properties of the Almon lag technique when one searches for degree of polynomial and lag, Jornal of the American Statistical Association, 70, 606-612, (1975).

[6] P. Schmidt, R.N. Wauld, The Almon lag technique and the monetary versus fiscal policy debate, Journal of the American Statistical Association, 68, 11-19,(1973).

[7] S. Almon, The distributed lag between capital appropriations and expenditures, Econometrica, 33, 178-196, (1965).

[8] P. Schmidt, Sicles, On the efficiency of the Almon lag technique, International Economic Review, (1975).

[9] R. Tadeusiewicz, Sygnał mowy, Warszawa, 1988.

[10] W. Wilnier, L. Pieszies, Oczierki po biołogiczeskoj kibiernietikie, Mińsk, Wyzszaja Szkoła, 1977.

[11] J. Imiołczyk, Prawdopodobieństwo subiektywne wyrazów podstawowy słownik frekwencyjny języka polskiego, PWN, Warszawa-Poznań, 1987.

[12] W. M. Głuszkow, Encykłopedia kibiernietiki, Gławnaja Redakcja Ukrainskoj Sowieckoj Encykłopedii,tom wtoroj, Kijew, 1975.