

Antoni Leon Dawidowicz
Uniwersytet Jagielloński, Kraków
Antoni.Leon.Dawidowicz@im.uj.edu.pl

WSPÓŁPRACA WITOLDA MAŃCZAKA I HUGONA STEINHAUSA W DZIEDZINIE ZASTOSOWAŃ METOD MATEMATYCZNYCH W JĘZYKOZNAWSTWIE

Słowa kluczowe: Hugo Steinhaus, Witold Mańczak, metody matematyczne w lingwistyce
Keywords: Hugo Steinhaus, Witold Mańczak, mathematical methods in linguistics

Wstęp

Przedmiotem niniejszego artykułu jest treść listów Hugona Steinhausa do Witolda Mańczaka dotyczących propozycji zastosowania metod matematycznych do pewnych problemów językoznawstwa. List z roku 1957 jest propozycją adaptacji na grunt językoznawstwa pewnej powszechnie już wtedy stosowanej metody. Ciekawsze są natomiast listy z roku 1958. Zawierają one propozycję pewnej modyfikacji taksonomii wrocławskiej (Florek et al. 1951a, b) umożliwiającą jej zastosowanie w zagadnieniach klasyfikacji języków. Zaproponowana metoda dendrytu odwrotnego nie jest nigdzie opublikowana.

1. Rozkład Poissona

Jednym z podstawowych twierdzeń rachunku prawdopodobieństwa jest prawo małych liczb Poissona. Mówi ono, z grubsza biorąc, że jeżeli mamy zmienną losową przyjmującą wartości naturalne, o której nic poza tym nie wiemy, możemy przyjąć,

że ma ona rozkład Poissona, tzn. prawdopodobieństwo, że przyjmie ona wartość k , jest równe

$$\frac{\lambda^k}{k!} e^{-\lambda},$$

gdzie λ jest średnią wartością tej zmiennej¹. Prawo to empirycznie weryfikował np. Bortkiewicz (Bortkewitsch 1898), analizując liczbę śmiertelnych kopnięć żołnierzy przez konie w kawalerii pruskiej i na podstawie danych zebranych w ciągu 20 lat z 14 korpusów kawalerii wykazał, że jest ona zgodna z rozkładem Poissona (błąd jest na trzecim miejscu po przecinku).

W liście z 8 sierpnia 1957 r. Hugo Steinhaus proponuje Witoldowi Mańczakowi algorytm analizy częstości występowania sufiksu *-ak* w rzeczownikach. Propozycja procedowania jest następująca. Należy znaleźć w tekście pierwszy *-ak* i, poczynając od niego, podzielić tekst na odcinki o tej samej długości, rozumianej jako ilość uderzeń w klawisze maszyny do pisanja. W każdym segmencie należy obliczyć ilość *-ak*-ów i oznaczyć odpowiednio przez n_i liczbę fiszek z *i* *-ak*-ami (w liście jest założone, że $i = 1, \dots, 7$, ale nie jest to istotne z punktu widzenia rozumowania). Tych fiszek powinno być

$$M \frac{c^k}{k!} e^{-c}.$$

We wzorze tym M oznacza globalną ilość fiszek, gdzie mogłyby występować *-ak*-i. Jest ona nieznana, gdyż rozpoczynamy przeszukiwanie od pierwszego *-ak* i nie wiemy, czy we wcześniejszym tekście *-ak* teoretycznie może występować. Innymi słowy, nie znamy liczby n_0 pustych kart. Można więc sformułować układ równań (zachowując oznaczenia z listu):

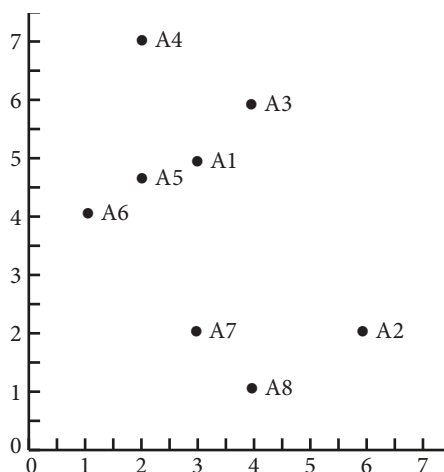
$$\left. \begin{aligned} Me^{-c} \cdot \frac{c^0}{1!} &= n_0 \\ Me^{-c} \cdot \frac{c^1}{1!} &= n_1 \\ Me^{-c} \cdot \frac{c^2}{2!} &= n_2 \\ \dots &\dots \\ Me^{-c} \cdot \frac{c^7}{7!} &= n_7 \end{aligned} \right\}.$$

¹ Symbol $k!$ (k -silnia) oznacza iloczyn wszystkich liczb naturalnych od 1 do n . Liczba e jest pewną stałą matematyczną i jest równa w przybliżeniu 2,718...

Niewiadomymi w tym układzie są, rzecz jasna, c , n_0 i M . Do oszacowania tych wielkości wystarczy więc znać trzy równania. W liście zaznaczono, że mogą wystarczyć nawet dwa równania, a pozostałe mogą służyć do weryfikacji metody. Na zakończenie listu Steinhaus wspomina, że zastosował tę metodę do szacowania liczby poległych na wojnie w oparciu o klepsydry (Kopocińska, Kopociński 2007a, b).

2. Taksonomia wrocławska

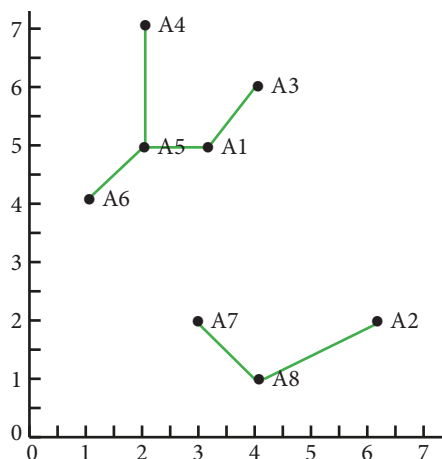
Dla dalszych rozważań omówimy teraz pewną technikę grupowania obiektów zwaną taksonomią wrocławską lub *cluster analysis*. Dla większej czytelności przedstawię tę procedurę na przykładzie. Wyobraźmy sobie, że na płaszczyźnie zadaliśmy zbiór punktów:



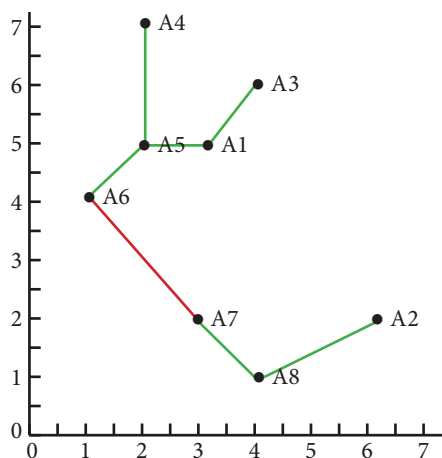
Konstruujemy tablicę, w której umieszczamy odległości między punktami (ze względów praktycznych umieszczam w tabeli kwadraty odległości).

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	18	2	5	1	5	9	17
A2	18	0	20	41	25	29	10	5
A3	2	20	0	5	5	13	17	25
A4	5	41	5	0	4	10	26	40
A5	1	25	5	4	0	2	10	20
A6	5	29	13	10	2	0	8	18
A7	9	10	17	26	10	8	0	2
A8	17	5	25	40	20	18	2	0

W każdym wierszu tej tablicy wybieramy odległość najmniejszą i zaznaczamy ją na naszym rysunku. W ten sposób otrzymujemy pewną ilość linii łączących zadane punkty:



Jeśli wszystkie punkty nie są ze sobą połączone, procedurę powtarzamy, traktując jako punkty poszczególne składowe spójne, a jako odległości między skupieniami – odległości między dwoma najbliższymi sobie punktami z każdego skupienia, i powtarzamy procedurę, aż wszystkie punkty są ze sobą połączone, czyli uzyskamy tzw. dendryt wrocławski:



Należy zwrócić uwagę, że figura, która powstaje, nie zawiera cykli, tzn. od każdego do każdego punktu można dojść tylko po jednej łamanej. Ustalając pewną wartość jako krytyczną, możemy usunąć wszystkie wiązania, których odległość jest

większa od krytycznej, i to, co zostaje, stanowi pogrupowanie obiektów takie, by w jednej grupie były najbardziej do siebie podobne.

Podany wyżej przykład dotyczył sytuacji, gdy klasyfikowane obiekty były punktami w przestrzeni euklidesowej. Można jednak rozważać innego rodzaju obiekty, byle była między nimi zdefiniowana odległość (metryka) rozumiana w następujący sposób:

Definicja 1. Niech dany będzie zbiór X . Odległością względnie metryką nazywamy odwzorowanie $\rho : X \times X \rightarrow R$ (tzn. odwzorowanie, które każdej parze punktów (x, y) przyporządkowuje liczbę $\rho(x, y)$) spełniające warunki:

- (1) $\rho(x, y) \geq 0, \rho(x, y) = 0$ wtedy i tylko wtedy, gdy $x = y$,
- (2) $\rho(x, y) = \rho(y, x)$,
- (3) $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$.

Wprowadzając odpowiednio zdefiniowane odległości, można metody taksonomii wrocławskiej zastosować do różnych sytuacji praktycznych. Gdy dana cecha wyraża się układem liczb, można potraktować je jako punkty odpowiedniej wielowymiarowej przestrzeni euklidesowej i zdefiniować odległość przez twierdzenie Pitagorasa. Zauważmy jednak, iż taksonomia wrocławska została opracowana dla potrzeb antropologii i dlatego potrzebne są nieraz inne definicje odległości. Najbardziej znanym przykładem odległości jest odległość Mahalanobisa definiowana następującym wzorem.

Niech $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_n)$. Ustalmy rodzinę r_{ij} spełniającą odpowiednie warunki. Wówczas

$$\rho(X, Y) = \sqrt{r_{11}(x_1 - y_1)^2 + r_{12}(x_1 - y_1)(x_2 - y_2) + \dots + r_{nn}(x_n - y_n)^2}$$

W przypadku, gdy r_{ij} jest równe 1 dla $i = j$ i 0 dla $i \neq j$ mamy do czynienia ze zwykłą odległością euklidesową. Stosowny dobór współczynników r_{ij} pozwala uwzględnić różny wpływ poszczególnych parametrów oraz ich zależność między sobą. W definicji odległości można również uwzględnić parametry dychotomiczne, czyli parametry, które są postaci „tak–nie”. Dzięki tak ogólnym możliwościom definiowania odległości dziś taksonomia wrocławska jest używana również w zagadnieniach ekonomii i rolnictwa (klasyfikacja odmian).

3. Dendryt odwrotny

Dendryt odwrotny to metoda uproszczenia procedury zaproponowana przez Steinhausa w listach do Witolda Mańczaka z 11 kwietnia i 19 października 1958 r. Problem jest najogólniej następujący. Chcąc dokonać klasyfikacji języków, czyli wyznaczyć odległości między nimi, musimy porównać dużo ich elementów (zjawisk gramatycznych, morfemów itp.). W liście zostało zaproponowane pewne rozwiązanie będące niejako odwróceniem obserwacji. Wybieramy kilka języków, o których wiemy, że stanowią jednorodną grupę (w tym przypadku są to języki romańskie), i określamy odległość między morfemami na tle wybranej grupy języków. Zacytujmy i skomentujmy teraz list.

Jeżeli morfem M_1 występuje a_1 razy w języku A , b_1 razy w B etc. ... zaś M_2 (występuje) a_2 [razy w języku] A , b_2 [razy w] B (etc.), to suma²

$$|a_1 - a_2| + |b_1 - b_2| + \dots |z_1 - z_2|$$

daje odległość morfemów $M_1 M_2$ na tle języków $A, B \dots Z$. W ten sposób znajdziemy wzajemne odległości morfemów $M_1, M_2 \dots M_{100}$ (jeżeli jest w ogóle tylko sto sensownych morfemów, to znaczy charakterystycznych *consensu ingeniorum*). Stąd powstanie dendryt morfemów. Mając go, znajdziemy jego punkty węzłowe, których będzie np. 20.

Innymi słowy, metodami taksonomii wrocławskiej dokonujemy klasyfikacji morfemów i do dalszych rozważań odległości między językami definiujemy już tylko przez porównanie tych wybranych morfemów. W ten sposób otrzymujemy o wiele prostsze potencjalne wzory na odległość między językami.

Ogólniejsze spojrzenie na dendryt odwrotny znaleźć możemy w liście z 19 października 1958 r. Zamiast morfemów porównuje się cechy języków. W liście jako przykłady są podane dwie cechy:

- Czy język ma neutrum?
- Czy 100 oznacza się przez „cent” (lub derywaty tego źródłosłowu)?

Należy zwrócić uwagę, że obie cechy są dychotomiczne, tzn. wyrażają się przez „tak-nie”. W liście podana jest przykładowa tabela dla hipotetycznych sześciu języków i obliczona jest odległość zdefiniowana następująco.

Ustalamy, że $a_i = 1$, jeżeli w i -tym języku występuje cecha A i $a_i = 0$, jeżeli nie występuje, analogicznie b_i jest równe zero lub jeden, w zależności od tego, czy występuje cecha B . Odległość między cechami A i B przy rozpatrywanych n językach wyraża się wtedy wzorem:

2 Wyrażenie $|x|$ oznacza bezwzględną wartość z x , czyli $|x| = x$ dla x dodatnich oraz $|x| = -x$ dla x ujemnych.

$$\rho(A, B) = \frac{|a_1 - b_1| + \dots + |a_n - b_n|}{a_1 + \dots + a_n + b_1 + \dots + b_n}.$$

Dla wyjaśnienia dalszej procedury zacytujmy znowu list:

Jaki jest cel tego? Otóż można uznać *a priori*, że języki romańskie są to indywidua jednakowo ważne, natomiast układy cech są mniej lub więcej arbitralne. Dlatego zaczynamy od języków i tworzymy dendryt iluś tam, np. 25 cech, co jest łatwe. Teraz patrząc na dendryt cech, wybierzemy z nich kluczowe, tzn. węzłowe, okaże się, że takich kluczowych jest 7. Te uznamy za najważniejsze i już przy tworzeniu Twoich dendrytów będziemy trzymać się tych 7 cech. Tak unikniemy arbitralnego doboru cech.

Z grubsza biorąc, postępujemy następująco:

- (1) Ustalamy *a priori* grupę kilku języków, która jest w miarę jednorodna.
- (2) Wybieramy listę cech, które mogą charakteryzować języki.
- (3) Na bazie ustalonego zbioru języków ustalamy odległości między cechami i tworzymy dendryt wrocławski (dendryt odwrotny).
- (4) W oparciu o dendryt odwrotny wybieramy cechy, które nazywamy kluczowymi.
- (5) W oparciu o te cechy kluczowe ustalamy odległości *m* między językami i tworzymy dendryt języków.

Zakończenie

Autorowi nie jest wiadome, czy metoda dendrytu odwrotnego została zastosowana zgodnie z sugestiami z listów. Tak czy owak, jest ona warta dokładniejszego zbadania pod kątem możliwości jej zastosowań.

Literatura

- VON BORTKEWITSCH L., 1898, *Das Gesetz der kleinen Zahlen*, „Monatshefte für Mathematik und Physik” t. 9, nr 1, s. A39–A41, B. G. Teubner, Leipzig.
- FLOREK K., ŁUKASIEWICZ J., PERKAL J., STEINHAUS H., ZUBRZYCKI S., 1951a, *Taksonomia wrocławska*, „Przegląd Antropologiczny” 17, s. 193–211.
- FLOREK K., ŁUKASIEWICZ J., PERKAL J., STEINHAUS H., ZUBRZYCKI S., 1951b, *Sur la liaison et la division des points d'un ensemble fini*, „Colloquium Mathematicum” 2, s. 282–285.
- KOPOCIŃSKA I., KOPOCIŃSKI B., 2007a, *Hugo Steinhaus problem of estimation of the war casualties on the base of contemporary press obituaries*, „Mathematica Applicanda” t. 35, nr 49/08, s. 155–161.
- KOPOCIŃSKA I., KOPOCIŃSKI B., 2007b, *Zagadnienie Steinhausa o szacowaniu strat wojennych na podstawie nekrologów prasowych*, „Matematyka Stosowana” 8, s. 155–161.

**Collaboration between Professors Witold Mańczak and Hugo Steinhaus
in the area of application of mathematical methods to linguistics**
Summary

This paper presents the letters written by Prof. Hugo Steinhaus to Prof. Witold Mańczak in which he proposes to apply certain mathematical methods to linguistic problems.

Kochany Wither!

[illegible]

W przyczynę metody:
Wzrost ograniczony jest i. l. materiału, w którym
 "ak" może występować. Roliny to j. u. Koimy namy
 tekstom szukać gdzie im jest podobna, ale nie rejestruje
 uściwale pmiennanego materiału, jeżeli w nim
 nie ma "ak". Jeżeli jednak w jakimś materiale
 znajduje się m. "ak" (m.p. "junak"), to koimy odliczyć
 tyle sztuk i pmiennego od tego miejsca 2000 sztuk
 pmiennego (t. u. 2000 sztuk w naszym i. pmiennego)
 i rejestrować ile rejestruje jest tam "ak" i. u. Tym
 sposobem otrzymujemy pewną liczbę N sztuk namy
 przyczyna: finka nr 375; karawia ko. Strain'sko, w. 1657
 Pomoi, str. 225 do 226, przyczyna: dworak, rejestruje 3.
 W końcu uzyskamy N sztuk, w których będzie n₁ sztuk
 z 1-m "akami" karta, n₂ sztuk z 2-m "akami" karta
 n₃ sztuk z 3-m i. t. d. ... n₇ sztuk z 7-m. Ten sposób

my tu widać Poissona udeży, którego w materiale
 1) obserwacji, jeżeli nadkied jest

$Me^{-c} \cdot \frac{c^k}{k!}$ obserwacji, w którym zjawisko występuje
 k razy

Stąd mamy równania

$$\left. \begin{aligned} Me^{-c} \cdot \frac{c^0}{0!} &= n_0 \\ Me^{-c} \cdot \frac{c^1}{1!} &= n_1 \\ Me^{-c} \cdot \frac{c^2}{2!} &= n_2 \\ Me^{-c} \cdot \frac{c^3}{3!} &= n_3 \end{aligned} \right\} \text{tried}$$

Jest to układ 3 równań o 3 niewiadomych.

M = uśredniona całkowita masa kani obijanych od
 2000 sztuk (M > N !!!), c średnia liczba, którą ma
 jedna kania, n_0 liczba pustych kani (które nie
 mamy, bo jej u nas nie ma: $M = N + n_0$). Pó-
 now jest znaczenie czegoś nie trzeba, co można unowli-
 wie kontrolę metali.

Mogłoby być założenie, że są tylko kania z 1 gum.
 z 2-gum ahami - to już wygenerować obliczenia
 frekwencji :

$$\left\{ \begin{aligned} Me^{-c} \cdot c &= n_1 \\ Me^{-c} \cdot \frac{c^2}{2} &= n_2 \end{aligned} \right\} \text{ z tego obliczamy c przez dzielnicę}$$

$$\left(\frac{c}{2} = \frac{n_2}{n_1} \right)$$

(Wzrost !!!) uśredniona wartość uśredniona do obliczenia
 strat uśredniona, kłopoty w gwarantach.
 Wnioskując w szczególności do Wro. Pawła uśredniona
 w Kuchni i robimy my o) i Treść.

Senkiewicz podawanie dla Ciebie

Adm. Kuchnia w
 nad Dunajcem,
 ul. J. Chr. 100,
 nad Północną w 556 u. n. H. Ch. 100

Hugo.

H.Steinhaus
Wrocław 12, ul. Orłowskiego 15

11 kwietnia, 1958 r.

Kochany Witku,

Dziękuję Ci za dawny list i za kartkę, która już ma adres nadawcy, ale nie ma daty - w każdym razie koło Wielkiej Nocy.

Interesuje mnie wynik dotyczący trzech języków romańskich, ale wciąż myślę nad zastosowaniem do lingwistyki mojej zasady dualizmu. Wyglądało by to tak:

Jeżeli morfem M_1 występuje a_1 razy w języku A, b_1 razy w B etc....
zaś " M_2 " a_2 " " " " b_2 " " " " ,
to suma $|a_1 - a_2| + |b_1 - b_2| + \dots + |z_1 - z_2|$ daje odległość morfemów $M_1 M_2$
na tle języków A, B, ..., Z. W ten sposób znajdziemy wzajemne odległości
morfemów M_1, M_2, \dots, M_{100} (jeżeli jest wogóle tylko sto sensownych morfem
mów, to znaczy charakterystycznych consensu ingeniorum). Stąd powstanie
dendryt morfemów. Mając go, znajdziemy jego punkty węzłowe, których będzie
n.p. 20. W nich stoją morfemy (n.p.) M_7, M_{13}, M_{18} itd. - te i tylko
te morfemy uznamy za ważne i na nich oprzemy najbliższą mapę (~~ś~~ dendryt)
języków A, B, ..., Z. - Nie jest to gotowa teoria, gdyż bez próby (hic Rhod-
us) trudno przewidzieć, co w trawie piszczy... dlatego chciałbym, że-
byś może na łatwym materiale spróbował, co tu można wygrać. -

Dziękuję Ci bardzo za pamięć i życzę pomyślnego i przyjemnego
pobytu w Paryżu. -

Tavj. H. H.

List Hugona Steinhausa z 11 IV 1958.

19.X.58

Ł. Henry Witkin,

Wzrostu Li chętniej, za list z 11/9. ~~1958~~
 Odpisuję dopiero teraz, bo cały mój czas był
 zajęty przygotowaniem odczytu i innych rzeczy;
 tego na 6 bulbr., a potem jesi list był
 aż chyba nie wstał w Paryżu. Odczyt trwa
 45 minut, przygotowanie 45 godzin.
 Oti i omówienie listowne sprawy dla syfikuji;
 isyhiu jest uienwile na i dlatego dłużej.
 Jęgi tu przyjechał na parę dni w liwi-
 pskie.

Dendryt odwrotny strumień się j.u. i
 Biene się jakkolwiek cechy isyhiu, n.p.
 taka, czy isyhiu ma centum albo n.p. taka
 ay w nim 100 oznaczają się przez "cent" (dub
 deruwały tego źródła słown). Biene się przez
 strumień isyhiu, n.p. isyhiu romaidio; ~~to~~ pro-
 uiedniemy, to jest id 6. Oznaczamy to isyhiu
 przez A B C D E F. Bzdriemy oznaczali przez
 przygotowanie cechy, przez 0 i j. bulbr. Na przy-

Przebieg:	A	B	C	D	E	F
neutrum	1	1	0	1	0	0
cent	0	1	1	1	0	1

Odległość "neutrum - cent" jest $\frac{3}{7}$, bo
 w trzech holennach jest różnica 1, a w syfiku-
 jedynek jest 7. Ułojąc odległości cod. mienimy
 budowa: dendryt cech (dojny na to 8 cech)
 na 6 isyhiu — będzie on odwrotny do

