Uniwersytet Wrocławski

CLARIN-PL

Struktury sekwencyjne w tekście

Prof. Adam Pawłowski

Instytut Podstaw Informatyki PAN, 12 czerwca 2017

PORZĄDEK WYKŁADU

1. Pojęcia (o czym mówimy)
2. Koncepcja badań
3. Formalizm (jedna z metod)
4. Badania empiryczne
5. Szeroka interpretacja

Uniwersytet Wrocławski

POJĘCIA: TEKST, SEKWENCYJNOŚĆ

Prof. Adam Pawłowski, IPI PAN, 12 maja 2017

POJĘCIA

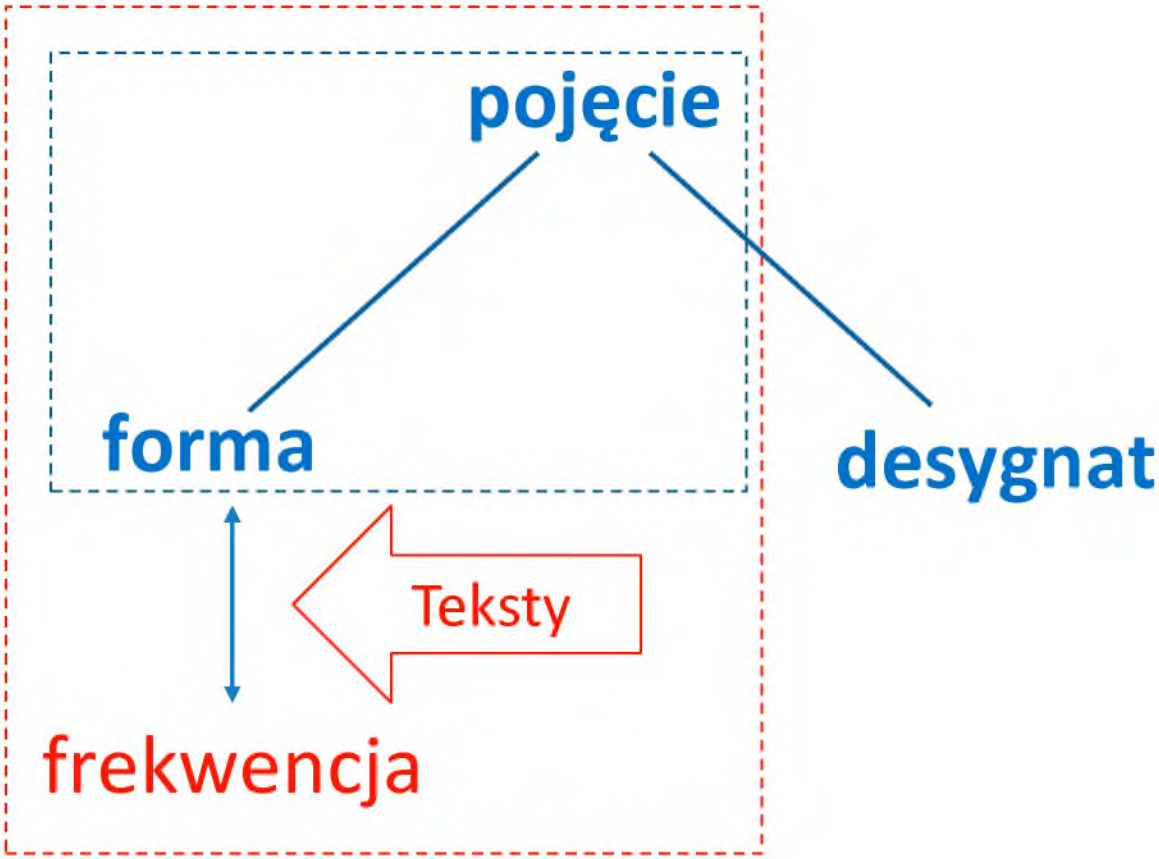
Tekst:

– ogólnie: ciąg wyrazów spełniający kryteria tekstowości (kohezja, koherencja, intencjonalność, akceptowalność, informatywność, sytuacyjność, intertekstowość)

– empirycznie: tekst jako sekwencja jednostek (liter/głosek, morfemów, wyrazów, zdań itd.)

– empirycznie: sekwencja odrębnych dokumentów lub tekstów wydzielonych, uporządkowanych wg jakiejś zasady

DEFINICJA ZNAKU JĘZYKOWEGO



SEKWENCYJNOŚĆ

Uporządkowanie elementów wg następstwa lub innej miary (np. czasu).

Dlaczego sekwencyjność jest ważna?

– relewantna i najbardziej oczywista cecha każdego wypowiedzenia lub tekstu języka

– cecha każdego wypowiedzenia lub tekstu języka wbudowana w schemat percepcji

SEKWENCYJNOŚĆ A INNE TERMINY

Tradycja:

relacje syntagmatyczne *vs* **paradygmatyczne**

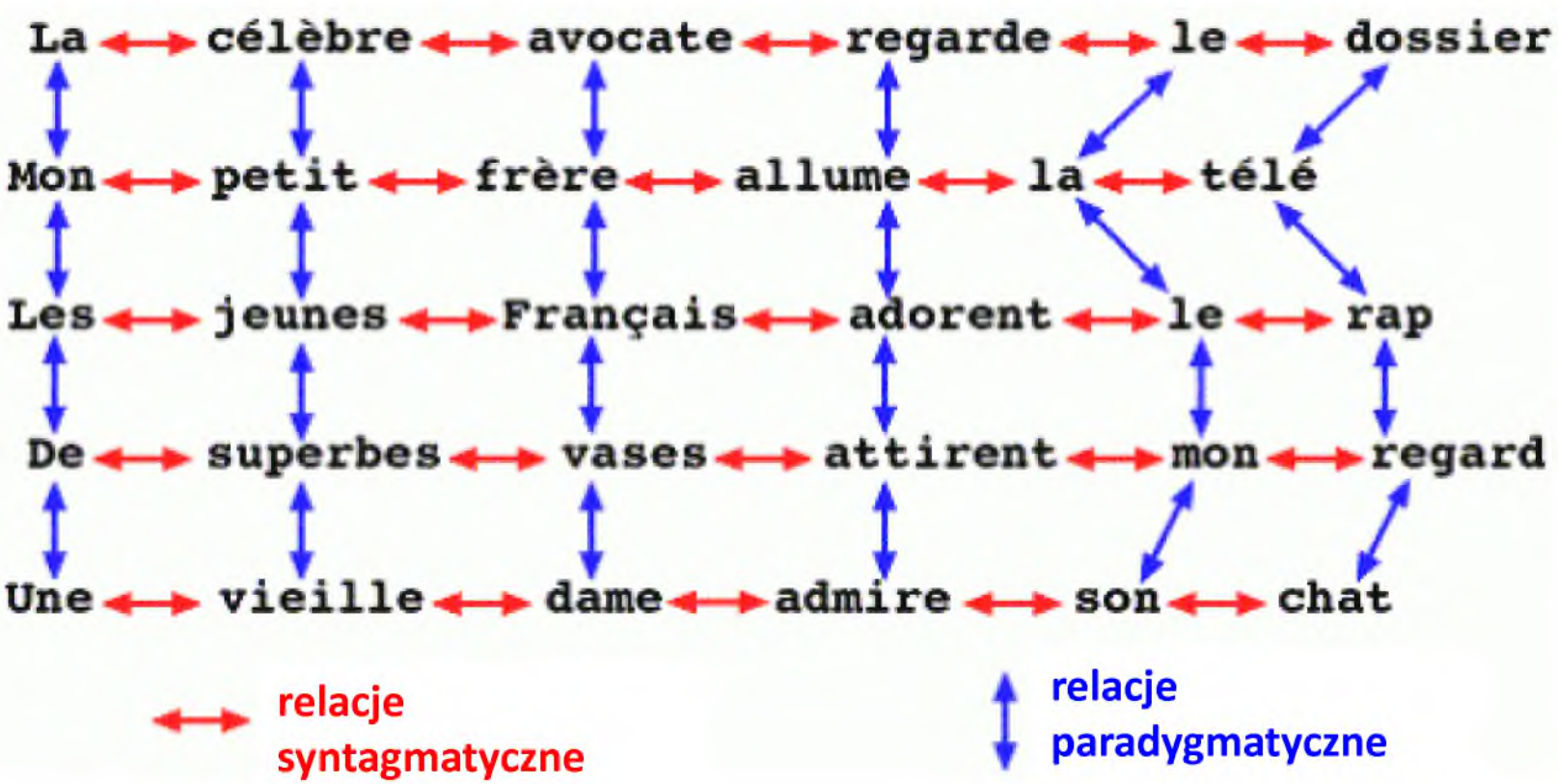
Lingwistyka kwantytatywna (G. Herdan):

language in the line *vs* **language in the mass**

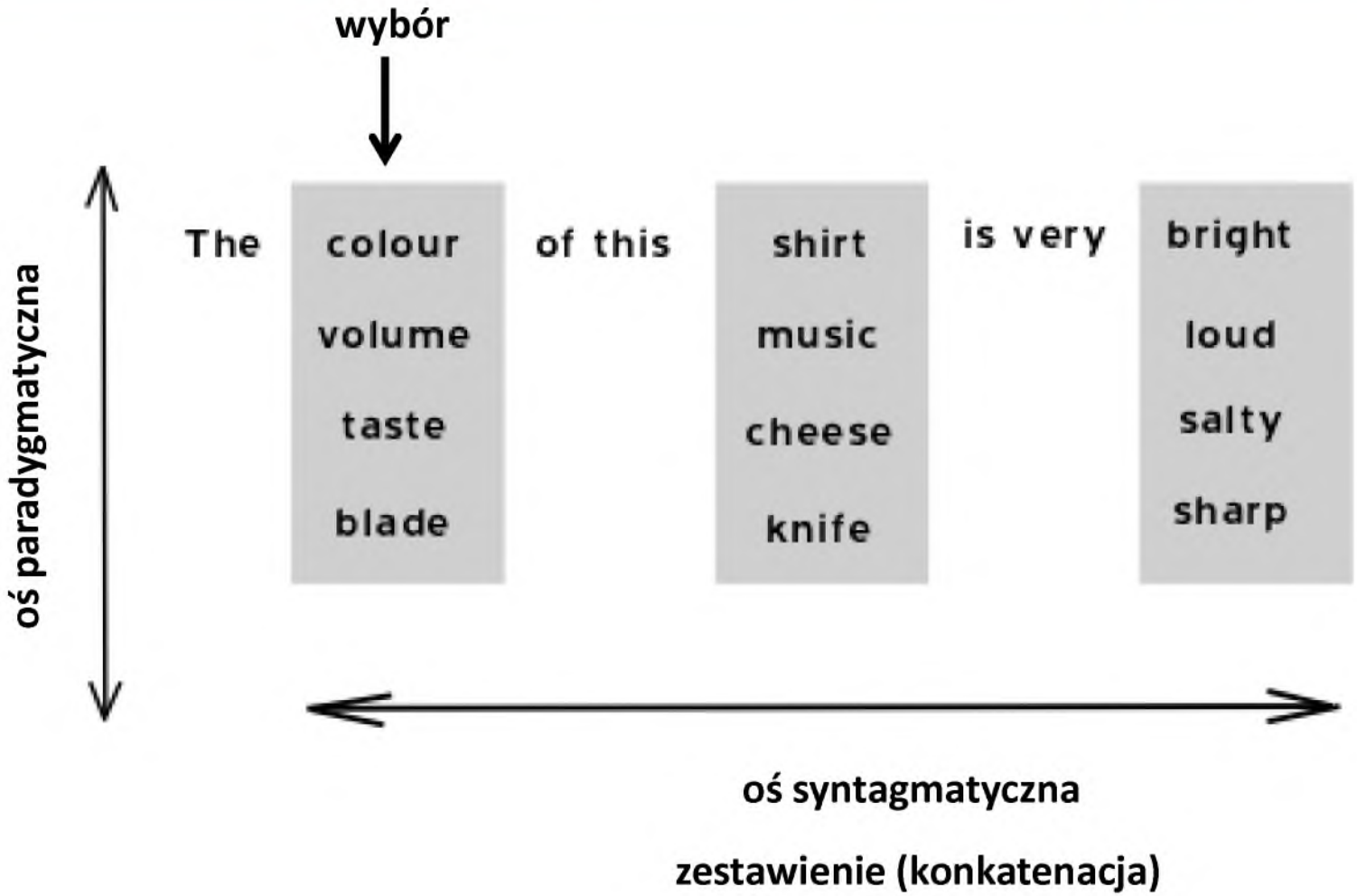
Przetwarzanie języka:

structured data *vs* **bag of words**

Relacje syntagmatyczne i paradygmatyczne



RELACJE SYNTAGMATYCZNE I PARADYGMATYCZNE



BAG OF WORDS *VS* STRUCTURE

Bag of words:

leksem 1 [frekwencja, parametr x]
leksem 2 [frekwencja, parametr x]
[...]
leksem *n* [frekwencja, parametr x]

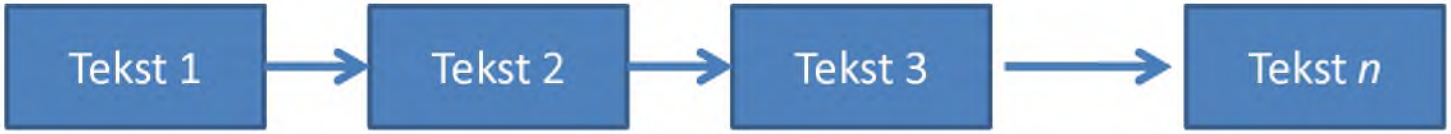
Struktura:

zdania, akapity, teksty,

SEKWENCYJNOŚĆ DOKUMENTÓW

To jest linia tekstu, reprezentująca sekwencję wyrazów w związku syntagmatycznym.

oś porządkowa



oś czasu lub oś porządkowa

Uniwersytet Wrocławski

Koncepcja badań

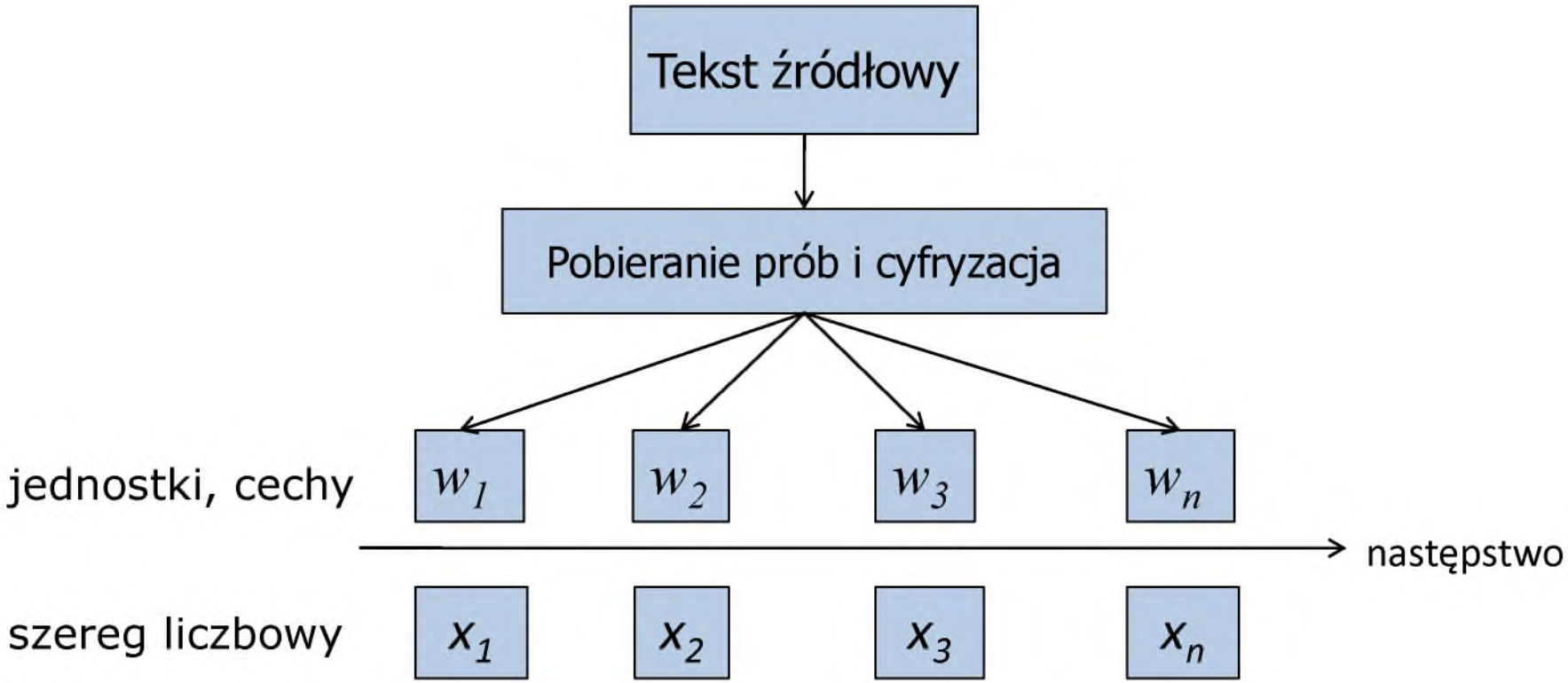
ISTOTA BADAŃ

Załóżmy, że w badanym języku występują dwa rodzaje jednostek określonych jako A i B, tworzących struktury liniowe („teksty”) w oparciu o pewne zasady składni. Załóżmy dalej, że dysponujemy następującym korpusem:

(1) AAAAAABBBBBB, (2) AAABBBAAABBB, (3) AABBAABBAABB, (4) ABABABABABAB, (5) AABBABABABAB, (6) BABBAABBBAAA

Widać gołym okiem, że zawsze F(A)=F(B)=5. Ale mimo to sekwencje są różne.

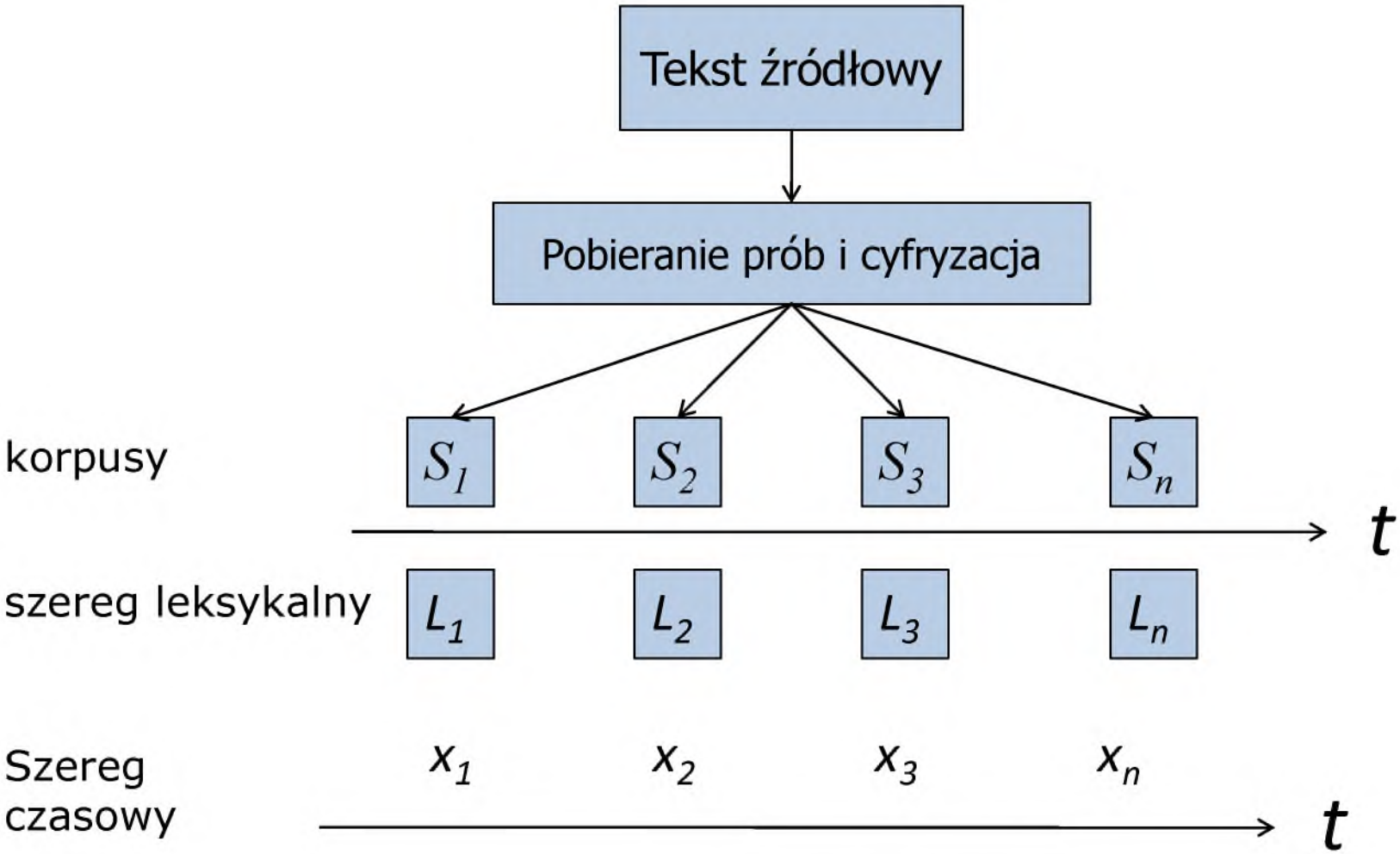
KONSTRUKCJA SZEREGU LICZBOWEGO



To jest linia tekstu, reprezentująca sekwencję wyrazów w związku syntagmatycznym.

[2, 4, 5, 6, 14, 9, 7, 1, 7,15]; [1, 1, 2, 2, 5, 3, 3, 1, 2,5]; etc.

KONSTRUKCJA SZEREGU CZASOWEGO



Jakie są szeregi leksykalne?

Stacjonarność

– szereg nie wychodzi poza pewne granice (nie "eksploduje");
– szereg nie ma wartości ujemnych (przed standaryzacją)

Ograniczona długość:

– nie ma źródła emitującego stale, są jednostkowe obserwacje;

Potrzeba interpretowalności:

– model czarnej skrzynki niewskazany, dane powinny mieć swoją interpretację;

WĄTPLIWOŚCI

Sekwencje w tekstach pisanych:

– zdania są osobnymi "jednostkami", a więc następstwo wyrazów ma ograniczony sens
– nakładają się porządki pisma (litera, wyraz) i mowy (sylaba, akcent)

Lepsze rozwiązanie:

– sekwencje w tekście mówionym (kategoria zdania trudna do zdefiniowania)

ALE

– nie da się przetwarzać długich sekwencji (literatury), potrzebne narzędzia do pomiarów, każde odczytanie może być inne

WĄTPLIWOŚCI

Sekwencje w seriach dokumentów/prób:

– założenie w ciemno, że proces istnieje;
– groźba myślenia życzeniowego (jeżeli proces nie istnieje, wymyślamy go)

Uniwersytet Wrocławski

FORMALIZM

METODY

Ekonometria:

– szeregi czasowe
– estymacja funkcji

Akustyka i telekomunikacja:

– modele w dziedzinie częstotliwości, analiza harmoniczna

Telekomunikacja:

– teoria informacji

Matematyka:

– procesy stochastyczne
– analizy *n*-gramów
– test serii

DLACZEGO AKURAT TEKSTY

Pierwsze badania sekwencyjne tekstu prowadził Andriej Andriejewicz Markow. Właśnie teksty (Puszkina, Lenina) były inspiracją i pierwszym materiałem teorii łańcuchów Markowa.

MARKOV A.A.

(1910), Recherches sur un cas remarquable d'épreuves dépendantes. In: *Acta Mathematica* 33, 87–104.

(1907), Issledovanie zamečateľnogo slučaja zavisimyh ispytanij. In: *Bulletin de l'Académie Impériale des Sciences* 6/1, 61–80.

(1911), Ob odnom slučae ispytanij, svjazannych v složnuju cep'. In: *Bulletin de l'Académie Impériale des Sciences* 6/5, 171–186.

(1913), Primer statističeskogo issledovanija nad tekstom „Evgenija Onegina", illjustrirujuščij svjaz' ispytanij v cep'. In: *Bulletin de l'Académie Impériale des Sciences* 6/7, 153–162.

1913. № 1.

ИЗВѢСТІЯ

ИМПЕРАТОРСКОЙ АКАДЕМІИ НАУКЪ.

VI СЕРІЯ.

15 ЯНВАРЯ.

BULLETIN

DE L'ACADÉMIE IMPÉRIALE DES SCIENCES

DE ST.-PÉTERSBOURG.

VI SÉRIE.

15 JANVIER.

№ 32720 / 1911 / Academy Museum

С.-ПЕТЕРБУРГЪ. — ST.-PÉTERSBOURG.

Примѣръ статистическаго изслѣдованія надъ текстомъ „Евгенія Онѣгина" иллюстрирующій связь испытаній въ цѣпь.

А. А. Марковъ.

(Доложено въ засѣданіи Физико-Математическаго Отдѣленія 23 января 1913 г.)

Наше изслѣдованіе относится къ послѣдовательности 20 000 русскихъ буквъ, не считая ъ и ь, въ романѣ А. С. Пушкина «Евгеній Онѣгинъ», которая заполняетъ всю первую главу и шестьнадцать строфъ второй.

Эта послѣдовательность доставляетъ намъ 20 000 связанныхъ испытаній, каждое изъ которыхъ даетъ гласную или согласную букву.

Соотвѣтственно этому мы допускаемъ существованіе неизвѣстной постоянной вѣроятности p буквѣ быть гласной и приближенную величину числа p нашемъ изъ наблюденій, считая число появившихся гласныхъ и согласныхъ буквъ. Кромѣ числа p мы найдемъ, также изъ наблюденій, приближенныя величины двухъ чиселъ p_1 и p_0 и четырехъ чиселъ $p_{1,1}$, $p_{1,0}$, $p_{0,1}$, $p_{0,0}$, представляющихъ такія вѣроятности: p_1 — гласной слѣдовать за гласной, p_0 — гласной слѣдовать за согласной, $p_{1,1}$ — гласной слѣдовать за двумя гласными, $p_{1,0}$ — гласной слѣдовать за согласной, которой предшествуетъ гласная, $p_{0,1}$ — гласной слѣдовать за гласной, которой предшествуетъ согласная и, наконецъ, $p_{0,0}$ — гласной слѣдовать за двумя согласными.

Эти обозначенія согласованы съ принятыми въ статьѣ моей «Объ одномъ случаѣ испытаній связанныхъ въ сложную цѣпь»; при ссылкѣ же на статью «Изслѣдованіе замѣчательнаго случая зависимыхъ испытаній» надо p_0 приравнять p_2. Противоположныя вѣроятности, буквѣ быть согласной, обозначимъ, какъ принято нами, буквою q съ тѣми же значками.

Разыскивая число p, мы находимъ для него сначала 200 приближенныхъ величинъ, изъ которыхъ затѣмъ выводимъ среднюю ариѳметическую.

А именно, мы разбиваемъ всю послѣдовательность 20 000 буквъ на 200 отдѣльныхъ послѣдовательностей по 100 буквъ и считаемъ, сколько гласныхъ въ каждой сотнѣ буквъ: мы получаемъ 200 чиселъ, которыя, по раздѣленіи на 100, даютъ 200 приближенныхъ величинъ p.

При счетѣ числа гласныхъ мы имѣемъ въ виду сохранить возможность образовать другія соединенія по 100 буквъ; каждую изъ нашихъ сотенъ мы располагаемъ въ квадратъ по десяти строкъ и десяти столбцовъ, сохраняя порядокъ буквъ:

1	2	3	4	5	6	7	8	9	10
11,	12,	13,	14,	15,	16,	17,	18,	19,	20
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
91,	92,	93,	94,	95,	96,	97,	98,	99,	100.

Считаемъ сколько гласныхъ въ каждомъ столбцѣ, въ отдѣльности, и соединяемъ числа по два:

$$1^{\text{го}} \text{ и } 6^{\text{го}},\quad 2^{\text{го}} \text{ и } 7^{\text{го}},\quad 3^{\text{го}} \text{ и } 8^{\text{го}},\quad 4^{\text{го}} \text{ и } 9^{\text{го}},\ 5^{\text{го}} \text{ и } 10^{\text{го}}.$$

Мы получаемъ такимъ образомъ для каждой сотни буквъ пять чиселъ, обозначаемыхъ нами символами

$$(1,6),\quad (2,7),\quad (3,8),\quad (4,9),\quad (5,10);$$

сумма ихъ

$$(1,6)+(2,7)+(3,8)+(4,9)+(5,10)$$

равна числу гласныхъ этой сотни.

Соединяя же по 500 буквъ вмѣстѣ, мы можемъ образовать новыя пять сотенъ буквъ: первую — изъ первыхъ и шестыхъ столбцовъ, вторую — изъ вторыхъ и седьмыхъ столбцовъ и т. д.

Число гласныхъ въ этихъ новыхъ сотняхъ опредѣляется, очевидно, суммами

$$\Sigma\,(1,6),\quad \Sigma\,(2,7),\quad \Sigma\,(3,8),\quad \Sigma\,(4,9),\quad \Sigma\,(5,10),$$

состоящими изъ соотвѣтствующихъ пяти слагаемыхъ.

Результаты нашего счета приведены въ сорока таблицахъ, каждая изъ которыхъ содержитъ: въ первой строкѣ — пять чиселъ (1,6) и ихъ сумму, во второй строкѣ — пять чиселъ (2,7) и ихъ сумму и т. д., а въ послѣдней строкѣ — число гласныхъ въ первой сотнѣ, во второй сотнѣ и т. д. и наконецъ число гласныхъ во всѣхъ пяти сотняхъ, уменьшенное для сбереженія мѣста на 200.

6 8 11 11 13 **49**	16 11 9 8 7 **51**	11 12 7 3 6 **42**	5 11 10 6 10 **42**	10 6 6 6 7 **35**
12 11 7 7 5 **42**	4 8 9 11 10 **42**	5 5 11 9 11 **41**	12 8 8 11 7 **46**	9 12 15 6 9 **51**
6 6 6 7 13 **38**	9 9 9 7 10 **41**	8 10 6 10 7 **41**	7 7 12 10 9 **45**	9 3 6 10 9 **37**
8 10 11 9 4 **42**	12 9 6 10 7 **44**	11 11 8 3 10 **43**	8 12 7 9 9 **45**	9 11 8 5 6 **39**
10 11 5 10 8 **44**	3 8 10 8 9 **38**	4 4 11 11 8 **41**	12 8 10 9 8 **47**	9 10 10 10 9 **48**
42 46 40 44 43 15	**44 45 43 44 43** 19	**42 42 43 39 42** 8	**44 46 47 45 43** 25	**46 42 45 37 40** 10
8 7 8 7 10 **40**	11 11 8 7 7 **44**	11 10 10 12 6 **49**	12 9 8 10 10 **49**	8 9 9 5 8 **39**
10 9 9 8 8 **44**	9 6 10 11 11 **47**	4 4 9 7 9 **33**	3 10 12 9 10 **44**	7 9 9 11 7 **43**
8 9 8 8 8 **41**	12 9 9 5 6 **41**	11 13 6 9 10 **49**	11 11 6 11 10 **49**	10 6 6 9 9 **40**
10 6 13 6 12 **47**	10 8 6 11 11 **46**	6 7 11 8 6 **38**	10 8 11 6 7 **42**	7 8 15 6 9 **45**
8 12 5 13 6 **44**	7 6 8 9 8 **38**	8 6 10 7 12 **43**	6 8 7 9 6 **36**	11 7 6 11 10 **45**
44 43 43 42 44 16	**49 40 44 43 43** 16	**40 40 46 43 43** 12	**42 46 44 45 43** 20	**43 39 45 42 43** 12
7 7 7 7 9 **37**	12 7 7 6 8 **40**	7 4 11 5 7 **34**	5 5 7 5 9 **31**	8 6 5 14 11 **44**
9 13 6 8 4 **40**	6 8 7 10 8 **39**	11 14 9 11 9 **54**	12 6 10 10 8 **46**	8 12 10 7 4 **41**
9 7 11 12 14 **53**	9 10 10 8 7 **44**	7 6 9 8 9 **39**	8 14 11 11 10 **54**	8 10 9 8 14 **49**
7 11 8 9 7 **42**	9 5 6 7 7 **34**	10 9 8 10 5 **42**	4 5 9 5 9 **30**	9 5 9 9 6 **38**
8 10 10 11 9 **48**	7 11 9 12 7 **47**	11 10 8 9 11 **49**	13 14 9 11 7 **54**	8 13 11 5 10 **47**
40 45 42 47 43 20	**43 41 39 44 37** 4	**46 43 45 43 41** 18	**42 42 46 42 43** 15	**41 46 44 43 45** 19
10 9 13 6 12 **50**	4 11 10 12 5 **42**	5 11 10 6 5 **37**	4 4 10 11 5 **34**	13 11 13 10 10 **57**
8 8 8 9 5 **38**	14 9 8 7 14 **52**	8 9 8 10 10 **45**	6 12 9 8 10 **45**	7 10 9 6 2 **34**
10 10 8 9 10 **47**	4 8 9 8 4 **33**	8 8 6 9 9 **40**	13 4 10 8 6 **41**	8 8 7 8 12 **43**
7 9 10 7 10 **43**	8 14 11 12 6 **51**	10 6 9 7 6 **38**	7 10 7 12 11 **47**	9 11 9 10 6 **45**
9 8 3 11 7 **38**	11 6 7 4 14 **42**	11 9 8 10 12 **50**	9 13 8 1 8 **39**	6 3 7 9 9 **34**
44 44 42 42 44 16	**41 48 45 43 43** 20	**42 43 41 42 42** 16	**39 43 44 40 40** 6	**43 43 45 43 39** 13
11 6 8 9 5 **39**	10 10 4 7 9 **40**	10 8 7 8 8 **41**	10 5 11 13 5 **42**	8 8 13 5 8 **42**
6 10 6 8 13 **43**	11 10 13 13 9 **56**	6 9 9 8 7 **39**	7 11 9 7 10 **44**	9 10 7 14 9 **49**
10 5 11 11 6 **43**	10 7 5 9 6 **37**	15 9 11 13 9 **57**	10 10 4 7 7 **38**	9 11 6 8 7 **41**
9 12 6 8 10 **45**	10 5 8 10 10 **43**	5 10 5 4 7 **31**	7 7 14 13 7 **48**	7 9 12 6 9 **43**
7 11 9 10 10 **47**	6 13 10 5 6 **40**	8 9 10 12 9 **48**	11 9 9 6 15 **50**	10 9 9 12 9 **49**
43 44 40 46 44 17	**47 45 40 44 40** 16	**44 45 42 45 40** 16	**45 40 47 46 44** 22	**43 47 47 45 42** 24
12 7 12 5 12 **48**	10 14 7 6 6 **43**	9 6 7 10 5 **37**	12 13 5 9 11 **50**	5 11 8 12 10 **46**
10 8 5 13 4 **40**	4 6 8 10 14 **42**	11 10 7 8 9 **45**	7 7 10 5 8 **37**	12 8 9 8 6 **43**
10 13 8 7 9 **47**	13 6 12 8 5 **44**	10 10 9 9 10 **48**	7 7 9 14 7 **44**	8 11 9 8 7 **43**
9 4 12 6 9 **40**	7 13 5 8 10 **43**	8 6 12 10 10 **46**	12 13 7 8 10 **50**	8 5 7 11 8 **39**
4 12 9 9 8 **42**	8 5 15 10 9 **47**	9 11 8 5 11 **44**	4 4 12 11 9 **40**	11 11 10 6 8 **46**
45 44 46 40 42 17	**42 44 47 42 44** 19	**47 43 43 42 45** 20	**42 44 45 47 45** 21	**44 46 43 45 39** 17
9 11 10 6 13 **49**	5 9 7 10 6 **37**	8 6 8 7 14 **43**	7 9 8 6 7 **37**	9 11 11 8 8 **47**
9 8 6 8 6 **37**	10 9 11 7 7 **44**	8 14 13 8 4 **47**	9 8 6 10 11 **44**	10 8 5 9 10 **42**
7 7 12 10 9 **45**	11 11 11 10 8 **51**	12 4 6 9 11 **42**	10 9 10 8 10 **47**	6 8 16 12 11 **53**
12 12 6 8 8 **46**	7 7 5 10 10 **39**	6 8 9 10 8 **41**	8 7 4 9 4 **32**	12 11 5 7 8 **43**
5 7 9 11 4 **36**	13 8 9 8 10 **48**	6 8 11 8 6 **39**	11 8 10 8 9 **46**	6 5 9 10 8 **38**
42 45 43 43 40 13	**46 41 43 45 41** 19	**40 40 47 42 43** 12	**45 41 38 41 41** 6	**43 43 46 46 45** 23
5 7 4 3 7 **26**	4 7 9 11 10 **41**	10 8 7 8 7 **40**	12 10 11 4 5 **42**	12 13 6 6 10 **47**
14 10 13 9 5 **51**	10 7 9 4 9 **39**	10 8 11 10 7 **46**	5 9 10 11 11 **46**	6 3 10 10 4 **33**
7 8 6 8 9 **38**	8 13 9 12 10 **52**	6 11 11 10 10 **48**	10 8 10 7 13 **48**	11 11 9 7 14 **52**
7 10 9 5 9 **40**	7 5 7 7 12 **38**	12 8 7 6 5 **38**	11 8 8 11 5 **43**	5 8 8 9 9 **39**
9 10 11 16 7 **53**	13 10 10 9 5 **47**	5 9 11 12 11 **48**	4 8 8 9 11 **40**	11 6 11 12 7 **47**
42 45 43 41 37 8	**42 42 44 43** 46 17	**43 44 47 46 40** 20	**42 43 47 42 45** 19	**45 41 44 44 41** 18

ANALIZA SZEREGÓW CZASOWYCH

Szeregiem czasowym nazywamy sekwencję obserwacji realizacji zmiennej losowej na osi czasu.

Wartości zmiennej mogą reprezentować obserwacje jednostkowe lub interwałowe (odcinki czasowe).

Obserwowana zmienna losowa może być reprezentacją (kwantyfikacją) dowolnej cechy badanego systemu lub obiektu. Wybiera się jednak cechy relewantne, czyli istotne z jakiegoś powodu.

ANALIZA SZEREGÓW CZASOWYCH

Szereg czasowy $X_t = \{x_1, x_2 \dots x_i\}$ może zawierać 3 składowe:

T – tendencja główna (trend)

P – oscylacje periodyczne

E – składowa losowa (szum)

Model szeregu zwykle ma postać addytywną:

$$TS = T + P + E$$

W badaniu szeregów leksykalnych wartości *TS* zawarte są w pewnym przedziale, dlatego szeregi takie w zasadzie są zawsze stacjonarne.

W modelu tym brakuje składowej "katastrofy", czyli jednorazowego, nagłego i gwałtownego zdarzenia.

ANALIZA SZEREGÓW CZASOWYCH

Każda składowa estymowana jest nieco inaczej:

- dla tendencji głównej funkcje monotoniczne
- dla oscylacji periodycznych f. periodyczne
- dla słabych składowych periodycznych modele szeregów typu autoregresji (AR), ruchomej średniej (MA) lub mieszane ARMA, SARMA, ARIMA
- dla składowej katastroficznej wystarczy analiza częstości

ANALIZA SZEREGÓW CZASOWYCH

Szereg czasowy X_t definiujemy jako:

$$X_t = \{x_1, x_2 \dots x_t\}$$

Średnią μ_x szeregu czasowego X_t jest wyrażenie:

(45) $$\mu_x = E(X_t)$$

którego estymatorem jest:

(46) $$m_x = \frac{1}{N}\sum_{t=1}^{N} x_t$$

gdzie N – długość szeregu
x_t – wartość szeregu w momencie lub na pozycji t

Wariancję szeregu czasowego X_t definiuje się jako:

(47) $$\sigma_x^2 = E(X_t - \mu_x)^2$$

Estymatorem wariancji (przy tych samych oznaczeniach) jest wielkość:

(48) $$s_x^2 = \frac{1}{N}\sum_{t=1}^{N}(x_t - m_x)^2$$

Autokowariancję szeregu czasowego X_t przy odstępie k definiujemy jako:

(49) $$\gamma_k = E\{(X_t - \mu_x)(X_{t+k} - \mu_x)\}$$

Estymatorem funkcji autokowariancji jest funkcja:

(50) $$c_k = \frac{1}{N-k}\sum_{t=1}^{N-k}(x_t - m_x)(x_{t+k} - m_x)$$

Autokorelację szeregu (ACF – ang. *autocorrelation function*) można określić jako unormowaną funkcję autokowariancji:

(51) $$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma_x^2}$$

a jej estymatorem jest funkcja:

(52) $$r_k = \frac{c_k}{c_0} = \frac{c_k}{s_x^2}$$

PODSTAWOWE TYPY SZEREGÓW

Szeregiem autoregresji rzędu p, oznaczanym AR(p), nazywać będziemy ciąg wartości x_t opisywanych modelem o postaci[80]:

(54) $$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_p x_{t-p} + e_t$$

gdzie a_i – współczynniki modelu
e_i – wartości losowe o rozkładzie $N(0,1)$

Szeregiem średniej ruchomej rzędu q, oznaczanym MA(q), nazywać będziemy ciąg wartości x_t opisywanych modelem o postaci:

(56) $$x_t = e_t - b_1 e_{t-1} - b_2 e_{t-2} - \ldots - b_q e_{t-q}$$

gdzie b_i – współczynniki modelu
e_i – wartości szeregu losowego o rozkładzie $N(0,1)$

PODSTAWOWE TYPY SZEREGÓW

Szeregiem mieszanym typu ARMA(p, q) nazywać będziemy ciąg wartości opisywanych następującym modelem:

(58) $$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \ldots + a_p x_{t-p} - b_1 e_{t-1} - b_2 e_{t-2} - \ldots - b_q e_{t-q} + e_t$$

gdzie a_i – współczynniki modelu AR
b_i – współczynniki modelu MA
e_i – wartości szeregu losowego o rozkładzie $N(0,1)$

Szereg określa się jako sezonowy SARMA(P, Q)$_s$, jeżeli opisujący go model ma postać:

(59) $$x_t = a_s x_{t-s} + a_{2s} x_{t-2s} + \ldots + a_{ps} x_{t-ps} - b_s e_{t-s} - b_{2s} e_{t-2s} - \ldots - b_{qs} e_{t-qs} + e_t$$

gdzie a_i – współczynniki modelu AR
b_i – współczynniki modelu MA
s – odstęp sezonowy
e_i – wartości szeregu losowego o rozkładzie $N(0,1)$

IDENTYFIKACJA MODELU

	ACF	PACF
AR(p)	wygasa	urywa się na odstępie p+1
MA(q)	urywa się na odstępie q+1	wygasa
ARMA(p, q)	wygasa	wygasa

WSPÓŁCZYNNIK RYTMIZACJI

Syntetyczna miara sekwencyjnego uporządkowania tekstu

$$V_s = 100\% \left(1 - \frac{s_r^2}{s_{obs}^2} \right)$$

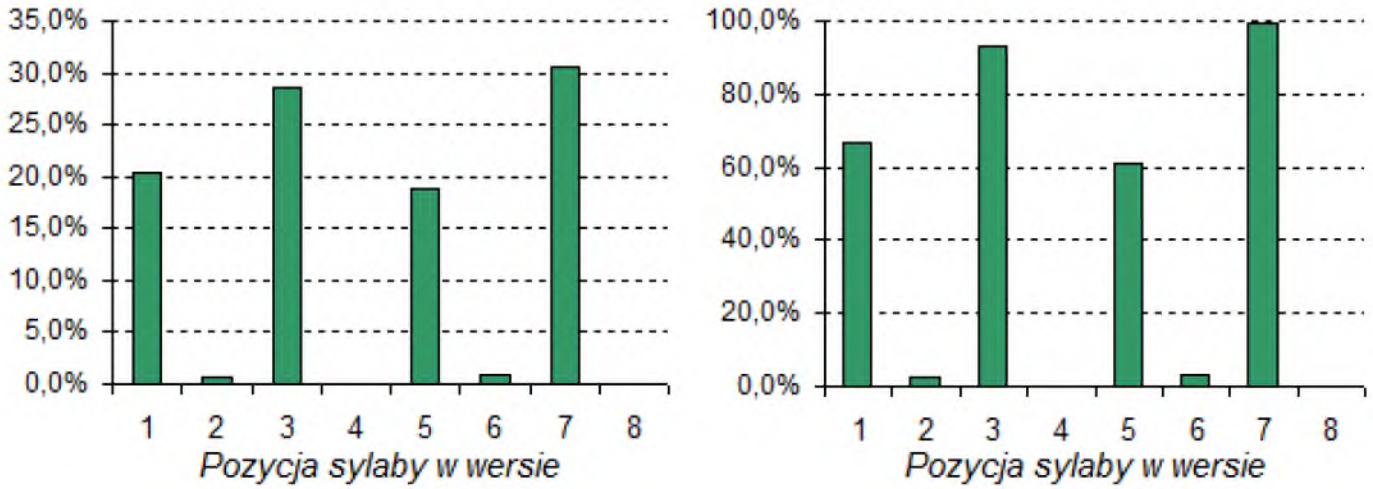
Uniwersytet Wrocławski

ZASTOSOWANIA

WIERSZ SYLABOTONICZNY

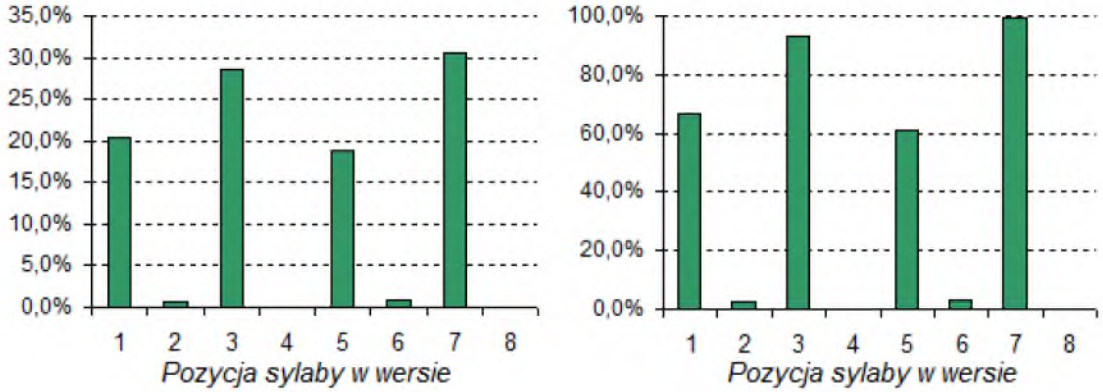
Każdy ptak i każde zwierzę	⊥	–	⊥	–	⊥	–	⊥	–
W swej komorze coś wybierze	–	–	⊥	–	⊥	–	⊥	–
I natychmiast to przyniesie	–	–	⊥	–	–	–	⊥	–
Jako udział w interesie	–	–	⊥	–	–	–	⊥	–

Rys. 6 Rozkład przycisków akcentowych w ośmiozgłoskowcu sylabotonicznym

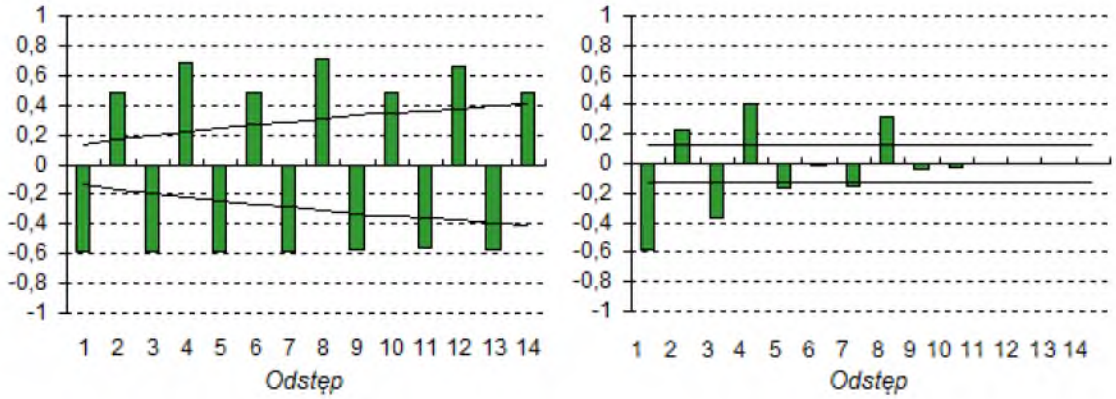


WIERSZ SYLABOTONICZNY

Rys. 6 Rozkład przycisków akcentowych w ośmiozgłoskowcu sylabotonicznym



Rys. 8 Autokorelacja (wykres lewy) i autokorelacja cząstkowa (wykres prawy) sekwencji przycisków akcentowych w wierszu sylabotonicznym[93]



WIERSZ SYLABOTONICZNY

Tab. 15 Identyfikacja modelu rytmiki wiersza sylabotonicznego

Typ modelu ($s_0^2 = 0{,}234$)	s_r^2	V_e	N
AR(1)	0,154	34%	1
AR(2)	0,147	37%	2
MA(1)	0,178	24%	1
ARMA(1,1)	0,104	55%	2
SARMA(1,1)(1,0)$_4$	0,098	58%	3
SARMA(1,1)(1,1)$_4$	0,088	62%	4
SARMA(1,1)(1,1)$_8$	0,082	65%	4

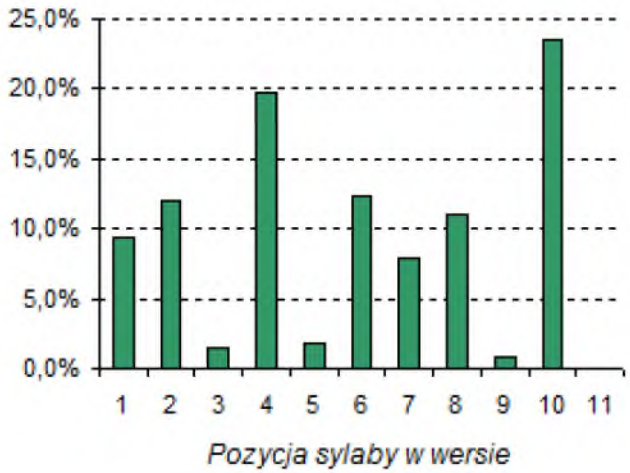
Oznaczenia:

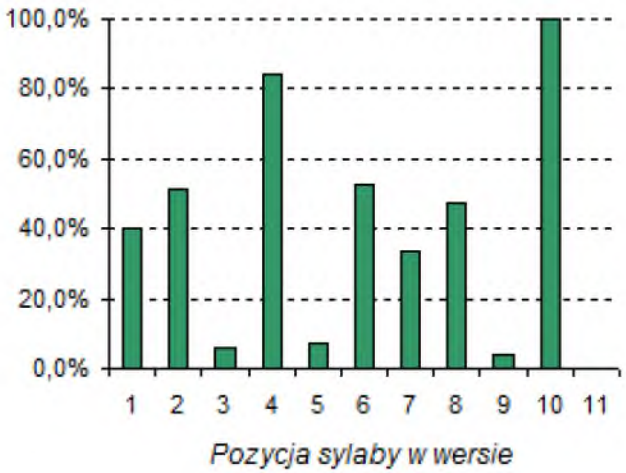
N – liczba parametrów modelu
s_0^2 – wariancja szeregu obserwowanego
s_r^2 – wariancja szeregu resztowego
V_e – procent wariancji wyjaśniony przez model (por. wzór 70)

$$(1+0{,}99B)(1-1{,}02B^8)x_t = (1+0{,}91B)(1-0{,}92B^8)e_t$$

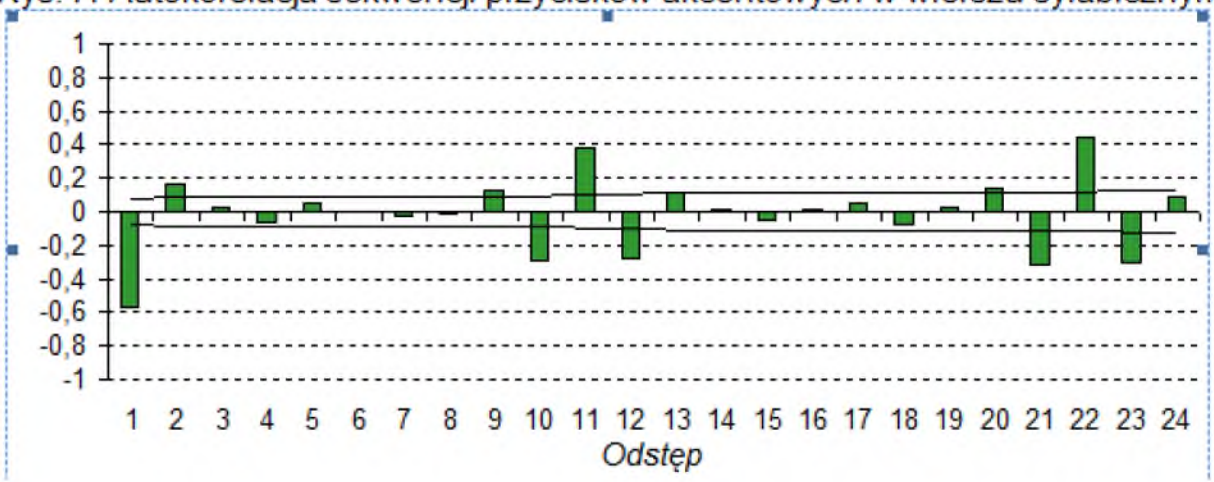
WIERSZ SYLABICZNY

Rys. 7 Rozkład przycisków akcentowych w jedenastozgłoskowcu sylabicznym





Rys. 11 Autokorelacja sekwencji przycisków akcentowych w wierszu sylabicznym



WIERSZ SYLABICZNY

Za panowania króla Stanisława
Mieszkał ubogi szlachcic na Podolu;
Wysoko potém go wyniosła sława;
Szczęścia miał mało w życiu, więcéj bolu;
Albowiem była to epoka krwawa
I kraj był cały na rumaku, w polu;
Łany, ogrody leżały odłogiem,
Zaraza stała u domu za progiem.

Tab. 16 Identyfikacja modelu rytmiki wiersza sylabicznego

Typ modelu ($s_0^2 = 0{,}239$)	s_r^2	V_e	N
MA(1)	0,162	32%	1
SAR(1)	0,204	15%	1
SMA(1)	0,218	9%	1
SARMA(0,0)(1,1)$_{11}$	0,157	34%	2
SARMA(0,1)(1,1)$_{11}$	0,123	48%	3

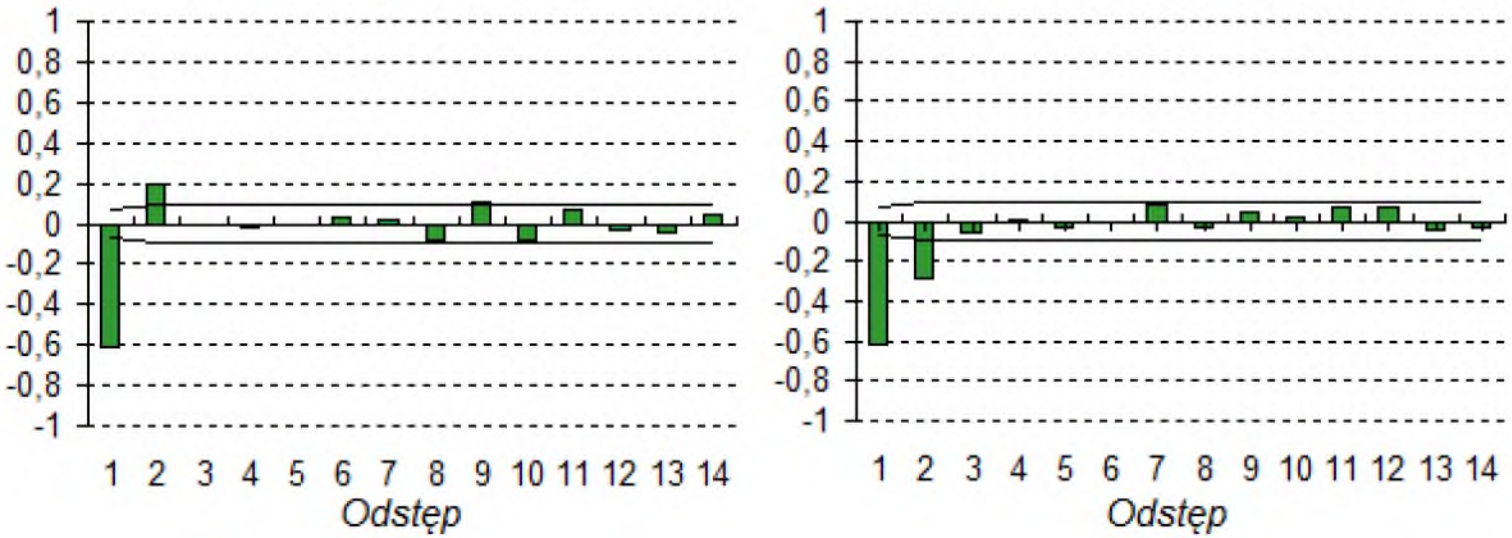
Oznaczenia:

N – liczba parametrów modelu
s_0^2 – wariancja szeregu obserwowanego
s_r^2 – wariancja szeregu resztowego
V_e – procent wariancji wyjaśniony przez model (por. wzór 70)

$$(1-0{,}99B^{11})x_t = (1-0{,}48B)(1-0{,}91B^{11})e_t$$

DYSKURS ORATORSKI

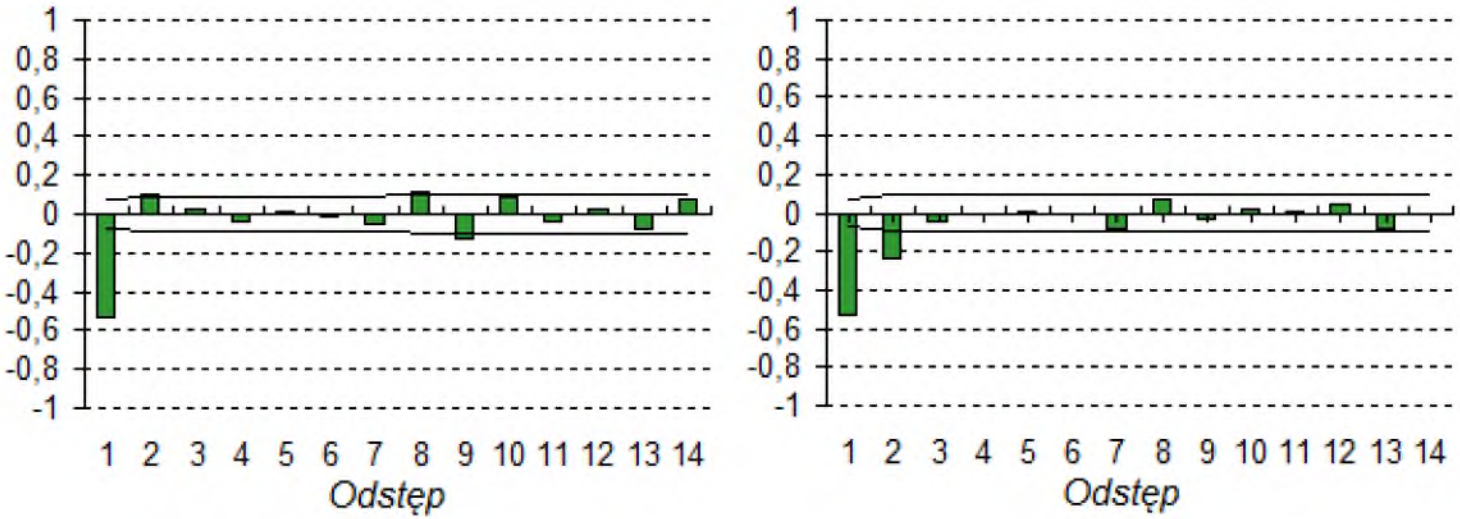
Rys. 17 Autokorelacja (wykres lewy) i autokorelacja cząstkowa (wykres prawy) sekwencji przycisków akcentowych w dyskursie oratorskim



$$x_t = (1 - 0{,}79B + 0{,}26B^2)e_t$$

PROZA ARTYSTYCZNA

Rys. 20 Autokorelacja i autokorelacja cząstkowa sekwencji przycisków akcentowych w prozie artystycznej



PORÓWNANIE

Tab. 17 Uporządkowanie rytmiczne stylistyczno-wersyfikacyjnych odmian polszczyzny

	Typ modelu	Próba 1	Próba 2	Próba 3	Średnia
Wiersz sylabotoniczny	$SARMA(1,1)(1,1)_8$	61%	65%	69%	65%
Wiersz sylabiczny	$SARMA(0,1)(1,1)_{11}$	45%	48%	51%	48%
Dyskurs oratorski	MA(2)	39%	43%	39%	40%
Proza artystyczna	MA(2)	36%	32%	33%	34%

WERYFIKACJA

Tab. 18 Liczby sylab akcentowanych i nie akcentowanych w badanych próbach

Próba	n	n_0	n_1	s^2	s^2/n	m_1
Wiersz sylabotoniczny 1	312	188	124	0,240	0,0008	0,3974
Wiersz sylabotoniczny 2	240	151	89	0,234	0,0010	0,3708
Wiersz sylabotoniczny 3	264	141	123	0,250	0,0009	0,4659
Wiersz sylabotoniczny	816	480	336	0,242	0,0003	0,4118
Wiersz sylabiczny 1	792	506	286	0,231	0,0003	0,3611
Wiersz sylabiczny 2	792	478	314	0,240	0,0003	0,3965
Wiersz sylabiczny 3	792	471	321	0,241	0,0003	0,4053
Wiersz sylabiczny	2376	1455	921	0,237	0,0001	0,3876
Dyskurs oratorski 1	731	465	266	0,221	0,0003	0,3639
Dyskurs oratorski 2	753	464	289	0,225	0,0003	0,3838
Dyskurs oratorski 3	668	405	263	0,224	0,0003	0,3937
Dyskurs oratorski	2152	1334	818	0,223	0,0001	0,3801
Proza artystyczna 1	727	461	266	0,227	0,0003	0,3659
Proza artystyczna 2	703	449	254	0,230	0,0003	0,3613
Proza artystyczna 3	531	339	192	0,228	0,0004	0,3616
Proza artystyczna	1961	1249	712	0,228	0,0001	0,3631

Oznaczenia:

n – wielkość próby
n_0 – liczba sylab nie akcentowanych
n_1 – liczba sylab akcentowanych
s^2 – wariancja z próby
m_1 – średnia liczba sylab akcentowanych

WERYFIKACJA

W celu porównania średnich z prób buduje się statystykę U o postaci:

$$U = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{77}$$

gdzie:

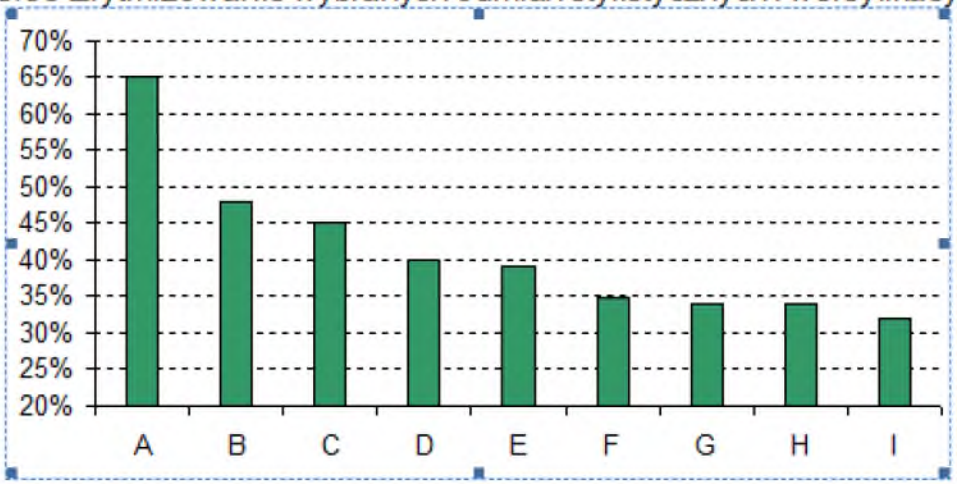
n_1, n_2 – wielkości porównywanych prób
$\bar{x}_1, \bar{x}_2$ – średnie z próby
s_1^2, s_2^2 – wariancje z próby

Tab. 20 Porównanie struktury rytmicznej wybranych odmian polszczyzny

	B	S	W	N
B		1,218	1,579	**2,411**
S	1,213		0,519	1,661
W	1,582	0,527		1,129
N	**2,396**	1,670	1,149	

PODSUMOWANIE

Rys.38 Zrytmizowanie wybranych odmian stylistycznych i wersyfikacyjnych polszczyzny



A - 65%
B - 48%
C - 45%
D - 40%
E - 39%
F - 35%
G - 34%
H - 34%
I - 32%

Oznaczenia:

A – wiersz sylabotoniczny 8-zgłoskowy (J.Brzechwa)
B – wiersz sylabiczny 11-zgłoskowy (J.Słowacki)
C – wiersz sylabiczny 13-zgłoskowy (A.Mickiewicz)
D – dyskurs oratorski (K.Wojtyła)
E – wiersz sylabiczny stroficzny (A.Puszkin)
F – proza artystyczna (M.Bułhakow)
G – proza artystyczna (I.Newerly)
H – proza artystyczna (J.Iwaszkiewicz)
I – styl prasowo-publicystyczny

INNE BADANIA TEKSTU

1) Porównanie języków o różnym typie prozodii
2) Porównanie tekstów a akcentem dynamicznym i iloczasowym
3) Porównanie języków o tendencji analitycznej i syntetycznej

Prof. Adam Pawłowski, IPI PAN, 12 maja 2017

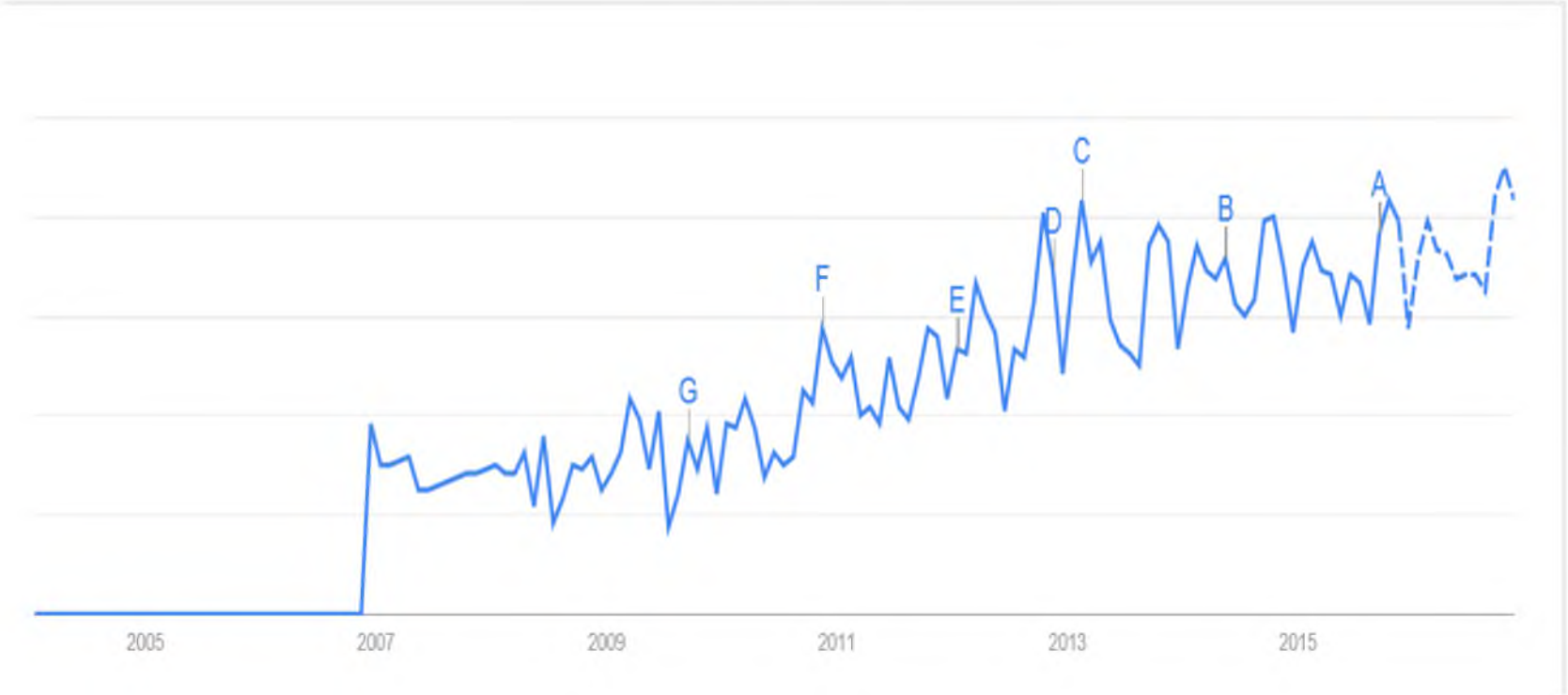
Uniwersytet Wrocławski

SZEREGI LEKSYKALNE

Prof. Adam Pawłowski, IPI PAN, 12 maja 2017

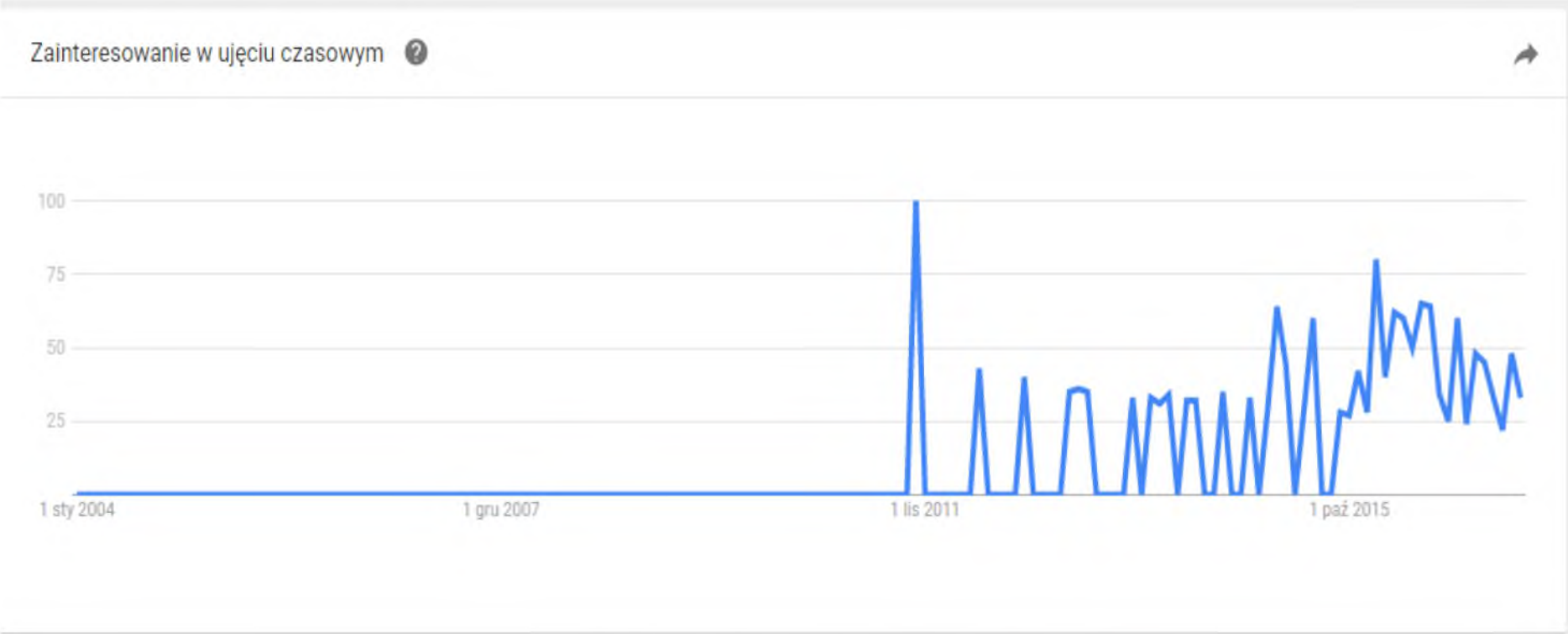
PRZYKŁAD: GENEZA HUMANISTYKI CYFROWEJ

Frekwencja frazy "digital humanities" w zasobach indeksowanych przez Google (dostęp 30.11.2015)

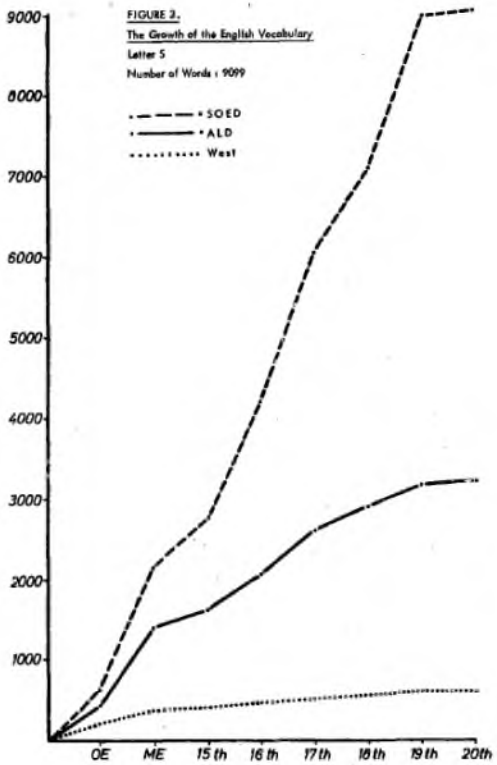


PRZYKŁAD: GENEZA HUMANISTYKI CYFROWEJ

Frekwencja frazy "humanistyka cyfrowa" w zasobach indeksowanych przez Google (dostęp 25.06.2017)



Czasy pionierów: przyrost słownictwa angielskiego



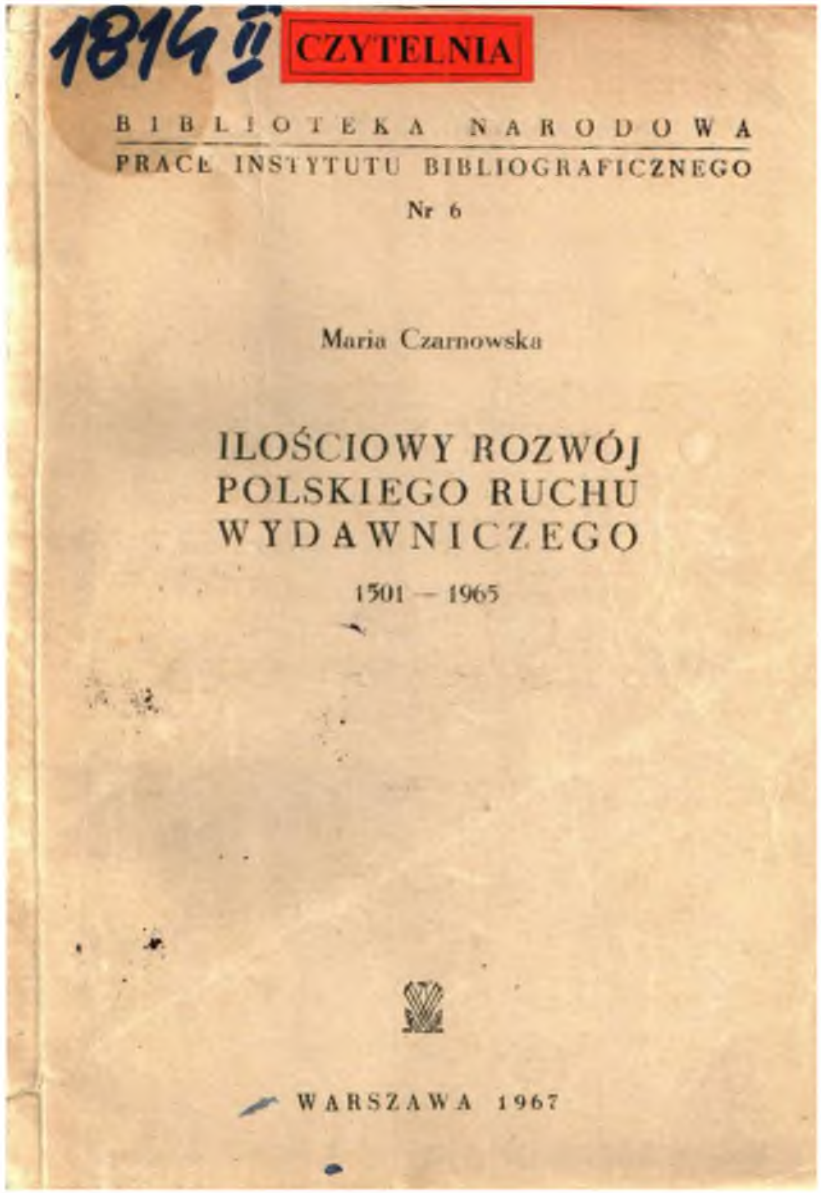
AN INVESTIGATION INTO THE VOCABULARY OF ENGLISH DICTIONARIES AND NEWSPAPERS
A Preliminary Report

Dieter WOLFF

Extrait de la Revue (R.E.L.O.)
III, 1 à 4, 1967. C.I.P.L. - Université de Liège - Tous droits réservés.

PRZYROST PRODUKCJI WYDAWNICZEJ

dr Maria Cecylia Czarnowska (1906–2001)


1814 II CZYTELNIA

BIBLIOTEKA NARODOWA
PRACE INSTYTUTU BIBLIOGRAFICZNEGO
Nr 6

Maria Czarnowska

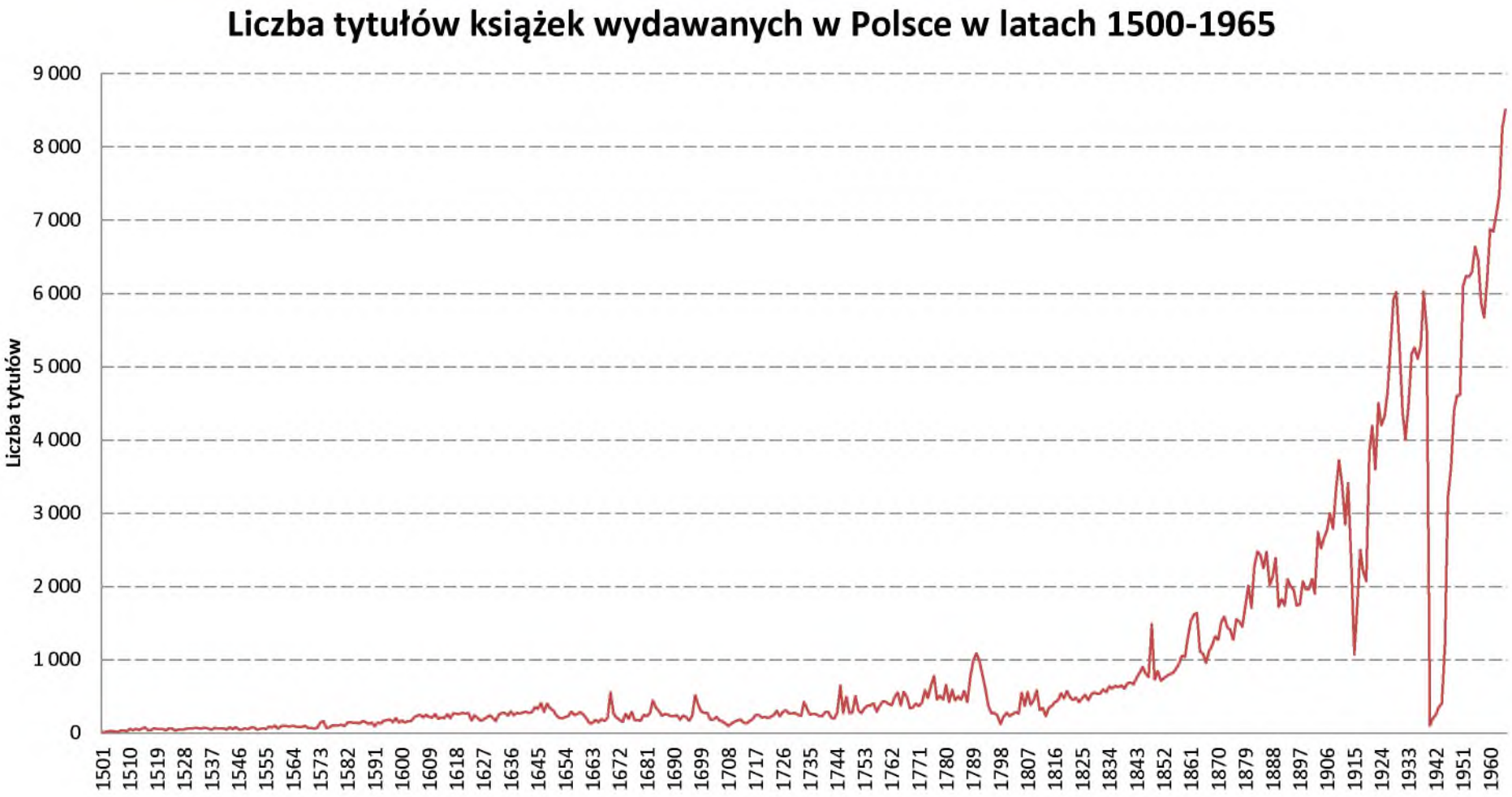
ILOŚCIOWY ROZWÓJ POLSKIEGO RUCHU WYDAWNICZEGO
1501 – 1965

WARSZAWA 1967

Badania chronologii słów (także wskazywanie słów miesiąca, roku etc.), to działania częściowo wpisujące się zakres **historii społecznej**.
Procesy zmian (trendy, cykle i katastrofy) można identyfikować stosując różne miary.

Uniwersytet Wrocławski

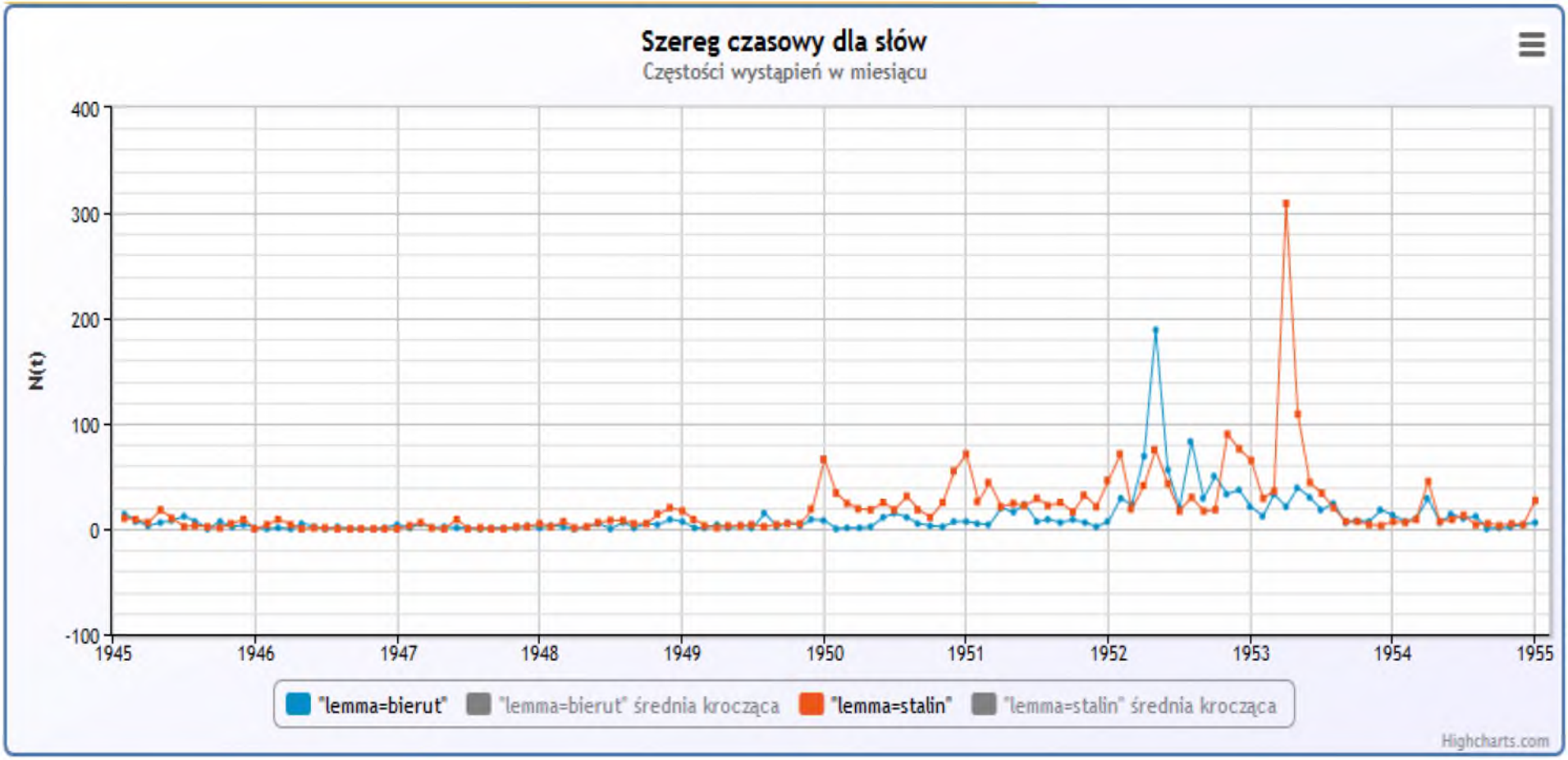
PRZYROST PRODUKCJI WYDAWNICZEJ (LICZBA TYTUŁÓW)



Źródło: M. Czarnowska (1967), *Ilościowy rozwój polskiego ruchu wydawniczego*. Warszawa: Biblioteka Narodowa. Opracowanie własne danych z *Aneksu*.

Szereg leksykalny z przesunięciem: nazwy miesięcy (1945-1954) (prasa polska, źródło **ChronoPress**)

Kiedy Predykcja jest bezsilna



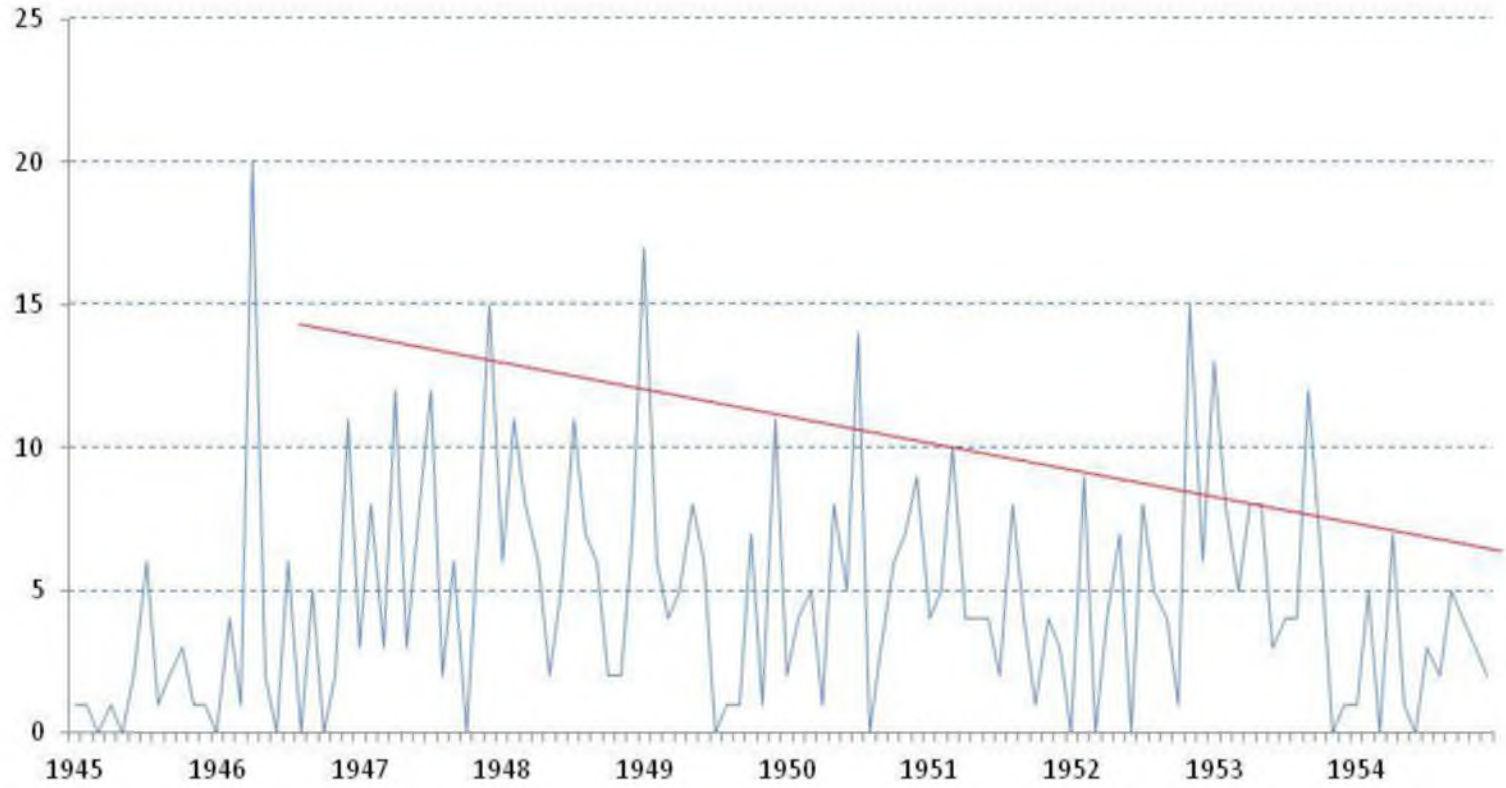
KIEDY PREDUKCJA JEST BEZSILNA

Graph these comma-separated phrases: terror ☐ case-insensitive

between 1800 and 2008 from the corpus English with smoothing of 0. Search lots of books

Badania poprawności politycznej. Szereg czasowy leksemu "Murzyn" (1945-1954)

(prasa polska, źródło **ChronoPress**)



Prof. Adam Pawłowski, IPI PAN, 12 maja 2017

Uniwersytet Wrocławski

CZYM JEST MODEL?

TRENDY – INTERESUJĄCY KOMENTARZ

Ted Kaczynski:

SOME PRINCIPLES OF HISTORY

99. Think of history as being the sum of two components: an erratic component that consists of unpredictable events that follow no discernible pattern, and a regular component that consists of long-term historical trends. Here we are concerned with the long-term trends.

100. FIRST PRINCIPLE. If a SMALL change is made that affects a long-term historical trend, then the effect of that change will almost always be transitory - the trend will soon revert to its original state. [...] If a small change in a long-term historical trend appears to be permanent, it is only because the change acts in the direction in which the trend is already moving, so that the trend is not altered but only pushed a step ahead.

TRENDY – INTERESUJĄCY KOMENTARZ

Ted Kaczynski:

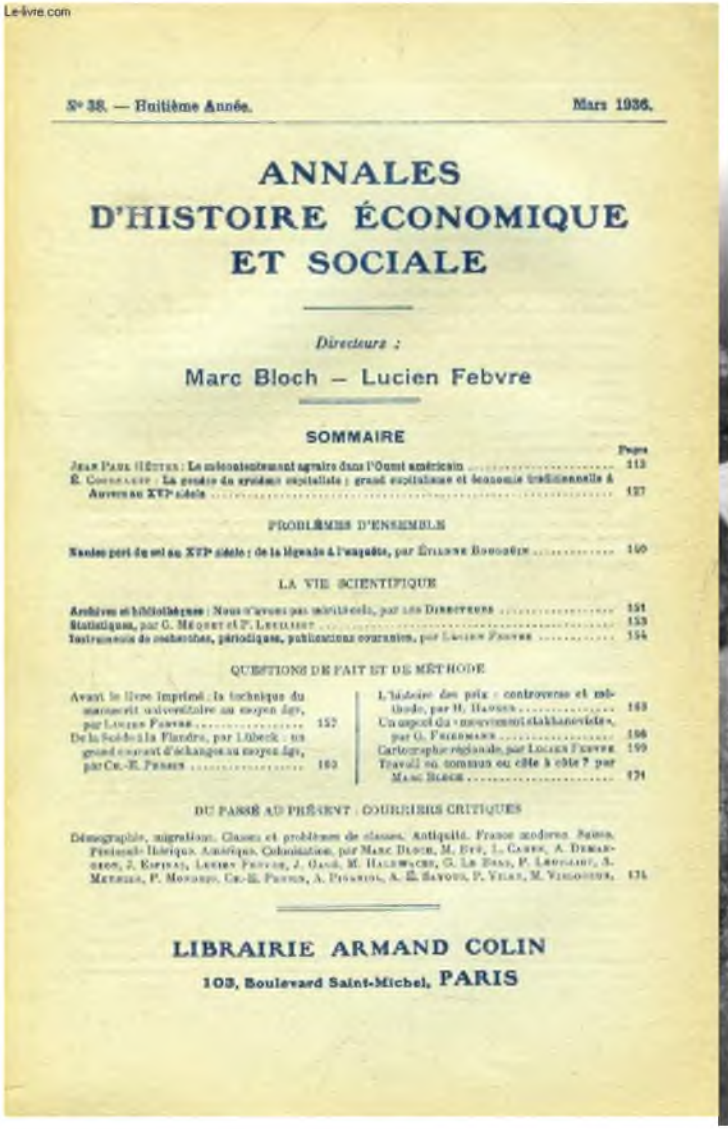
SOME PRINCIPLES OF HISTORY (cont.)

102. SECOND PRINCIPLE. If a change is made that is sufficiently large to alter permanently a long-term historical trend, than it will alter the society as a whole. In other words, a society is a system in which all parts are interrelated, and you can't permanently change any important part without change all the other parts as well.

103. THIRD PRINCIPLE. If a change is made that is large enough to alter permanently a long-term trend, then the consequences for the society as a whole cannot be predicted in advance. (Unless various other societies have passed through the same change and have all experienced the same consequences, in which case one can predict on empirical grounds that another society that passes through the same change will be like to experience similar consequences.)

104. FOURTH PRINCIPLE. A new kind of society cannot be designed on paper. That is, you cannot plan out a new form of society in advance, then set it up and expect it to function as it was designed to.

TRENDY – SZKOŁA ANNALES


Le-livre.com

N° 38. — Huitième Année. Mars 1936.

ANNALES D'HISTOIRE ÉCONOMIQUE ET SOCIALE

Directeurs :

Marc Bloch — Lucien Febvre

SOMMAIRE

Pages

Jean Paul Hüter : Le mécontentement agraire dans l'Ouest américain 113

E. Coornaert : La genèse du système capitaliste : grand capitalisme et économie traditionnelle à Anvers au XVIe siècle 127

PROBLÈMES D'ENSEMBLE

Naissance prix du sel au XVIe siècle : de la légende à l'enquête, par Étienne Bonnassie 140

LA VIE SCIENTIFIQUE

Archives et bibliothèques : Nous n'avons pas méritécoeli, par les Directeurs 151

Statistiques, par G. Mégret et P. Leuliet 153

Instruments de recherches, périodiques, publications courantes, par Lucien Febvre 154

QUESTIONS DE FAIT ET DE MÉTHODE

Avant le livre imprimé : la technique du manuscrit universitaire au moyen âge, par Lucien Febvre 157

De la Scelde à la Flandre, par Lübeck : un grand courant d'échanges au moyen âge, par Ch.-E. Perrin 163

L'histoire des prix : controverse et méthode, par H. Hauser 168

Un aspect du « mouvement stakhanoviste », par G. Friedmann 186

Cartographie régionale, par Lucien Febvre 199

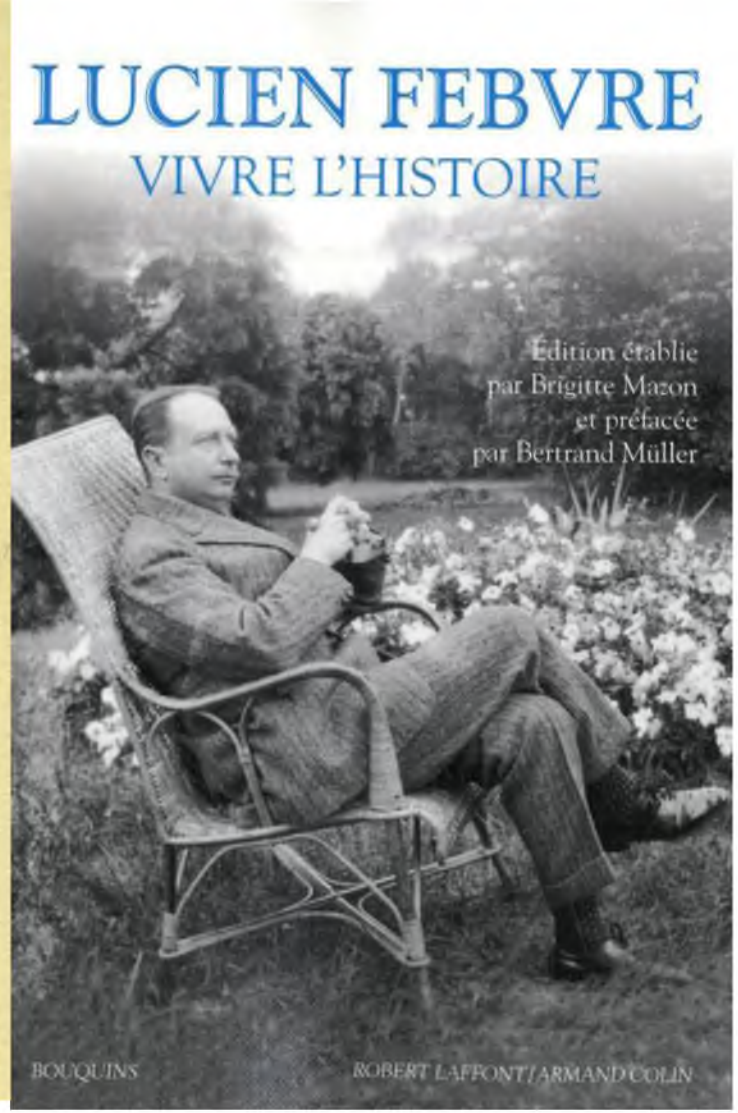
Travail en commun ou côte à côte ? par Marc Bloch 171

DU PASSÉ AU PRÉSENT : COURRIERS CRITIQUES

Démographie, migrations. Classes et problèmes de classes. Antiquité. France moderne. Pologne. Péninsule Ibérique. Amériques. Colonisation, par Marc Bloch, M. Eyé, L. Cahen, A. Demangeon, J. Espinas, Lucien Febvre, J. Gaus, M. Halbwachs, G. Le Bras, P. Lemoine, S. Meuriot, P. Monnier, Ch.-E. Perrin, A. Piganiol, A. E. Sayous, P. Vilar, M. Vignoussis. 174

LIBRAIRIE ARMAND COLIN

103, Boulevard Saint-Michel, PARIS



Interpretacja filozoficzna

Matematyka a rzeczywistość, dwa aspekty:

- epistemologiczny
- ontologiczny

Pierwotne systemy poglądów:

- platonizm
- arystotelizm
- pitagoreizm

INTERPRETACJA FILOZOFICZNA

Kierunki rozumowania:

- pragmatyczny (zadanie do wykonania)
- behawioralny (komfort dla umysłu)
- epistemologiczna koncepcja widza przed akwarium

Uniwersytet Wrocławski

DZIĘKUJĘ ZA UWAGĘ

Uwaga:
żaden fragment niniejszej prezentacji ekranowej (tekst, grafika, logotypy) nie może być powielany lub rozpowszechniany w żadnej formie i w żaden sposób bez uprzedniego zezwolenia jego twórcy. Wszelkie znaki graficzne, nazwy własne, logotypy i inne dane są chronione prawem autorskim i należą do ich właścicieli.