Uniwersytet Wrocławski

Modele sekwencji częstości leksemów w prasie na materiale korpusu ChronoPress

Prof. dr hab. Adam Pawłowski

Sympozjum „Słowa klucze”
Warszawa, 26 października 2015

Uniwersytet Wrocławski

Plan wykładu

1) Pojęcia

2) Główne założenia i opozycje

3) Modele sekwencji i koncepcje czasu:
   - model trendu monotonicznego;
   - model powtarzalnych oscylacji;
   - model katastrofy lub leksemu-komety;

4) Materiał testowy

5) Słowa miesiąca/roku a analiza chronologiczna

Kilka pojęć

Słowoforma (token)

Leksem (typ)

Częstość (słowoformy, leksemu)

Sekwencyjność i sekwencje w języku i tekście

Słowo klucz

Uniwersytet Wrocławski

Założenia i dylematy

1) Językoznawstwo wewnętrzne czy zewnętrzne?

1a) Relewantność językoznawstwa zewnętrznego i nieautonomiczność językoznawstwa.

1b) Nawiązanie do „społecznych" historii lub opisów zjawisk lub instytucji (*social history of...*).

2) Opozycja **jednostkowości** i **uniwersalności** w nauce o języku (obie kategorie uznaję za stopniowalne).

3) Opozycja **statyczności** i **dynamizmu** języka oraz jego naukowych reprezentacji.

4) Jedną z ram teoretycznych takich badań, ale z licznymi zastrzeżeniami, jest JOŚ.

Uniwersytet Wrocławski

Wzorce chronologiczne

Zmiany w świecie przedstawionym lub jego konstrukcji wyrażają się 4 modelami szeregów leksykalnych:

1. **Długotrwały ruch monotoniczny, trend** (czas linearny)
2. **Oscylacje periodyczne (**czas kolisty**)**
   - osadzone w cyklach natury (rytm pogody, rolnictwo: czas astronomiczny)
   - osadzone w kulturze (rytuały, obrzędy: czas antropologiczny)
   - osadzone w polityce i systemie prawa (czas polityczny)
3. **Schemat katastrofy lub leksemu-komety**
4. **Ruchy losowe**

   Tutaj „pozorne neosemantyzacje", czyli różne sensy tego samego leksemu, zależnie od tematu: przykład leksemy *wojna*, *miasto*).

*) Źródła: Pawłowski Adam (2006), Chronological analysis of textual data from the 'Wrocław Corpus of Polish'. *Poznań Studies in Contemporary Linguistics* (PSiCL) 41, 2006, 9-29. Pawłowski Adam (2008), *Chronological Corpus of Polish Press Texts*. In: M.A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, K. Trojanowski (red.), *Intelligent Information Systems* XVI: proceedings of the International IIS'08 Conference. Warszawa: Academic Publishing House EXIT, 481-486.

KONSTRUKCJA KATEGORII CZASU

1. **Czas astronomiczny** (cykle naturalne)

Leksemy związane z cyklicznymi zmianami pór roku i towarzyszącymi im zjawiskami. Przykładem są prace polowe, zjawiska pogodowe oraz pochodne (np. epidemie, akcje prewencyjne, katastrofy naturalne).

2. **Czas polityczny** (cykle polityczne i ekonomiczne)

Tutaj wybory (także te fasadowe), posiedzenia ciał kolegialnych, terminy uchwalania / ogłaszania aktów prawnych etc.

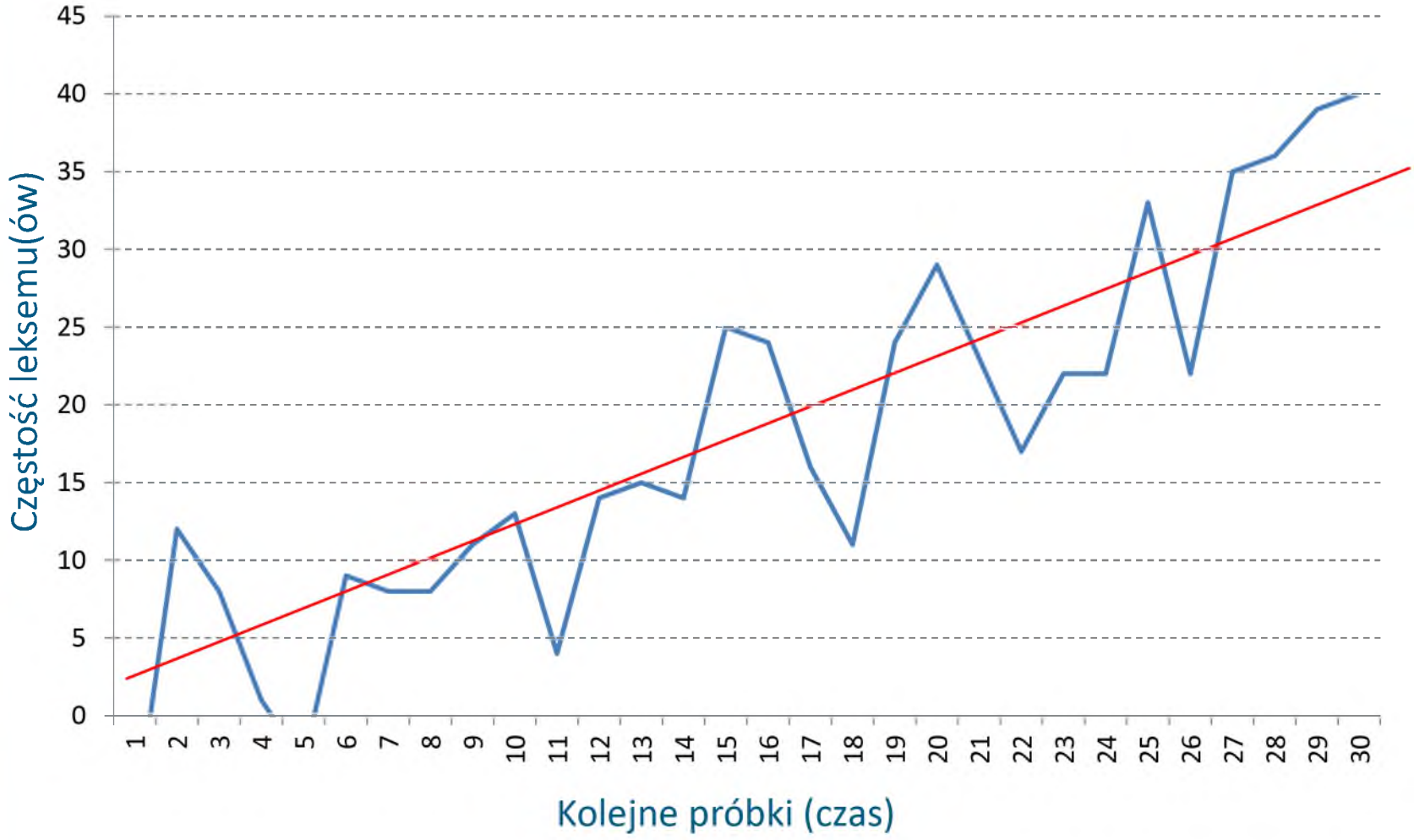
3. **Czas kulturowy** (cykle kulturowe)

Święta, rocznice, inne obrzędy lub powtarzające się rytuały kulturowe. Mogą mieć charakter oficjalny lub prywatny.

Uwagi na temat czasu kolistego

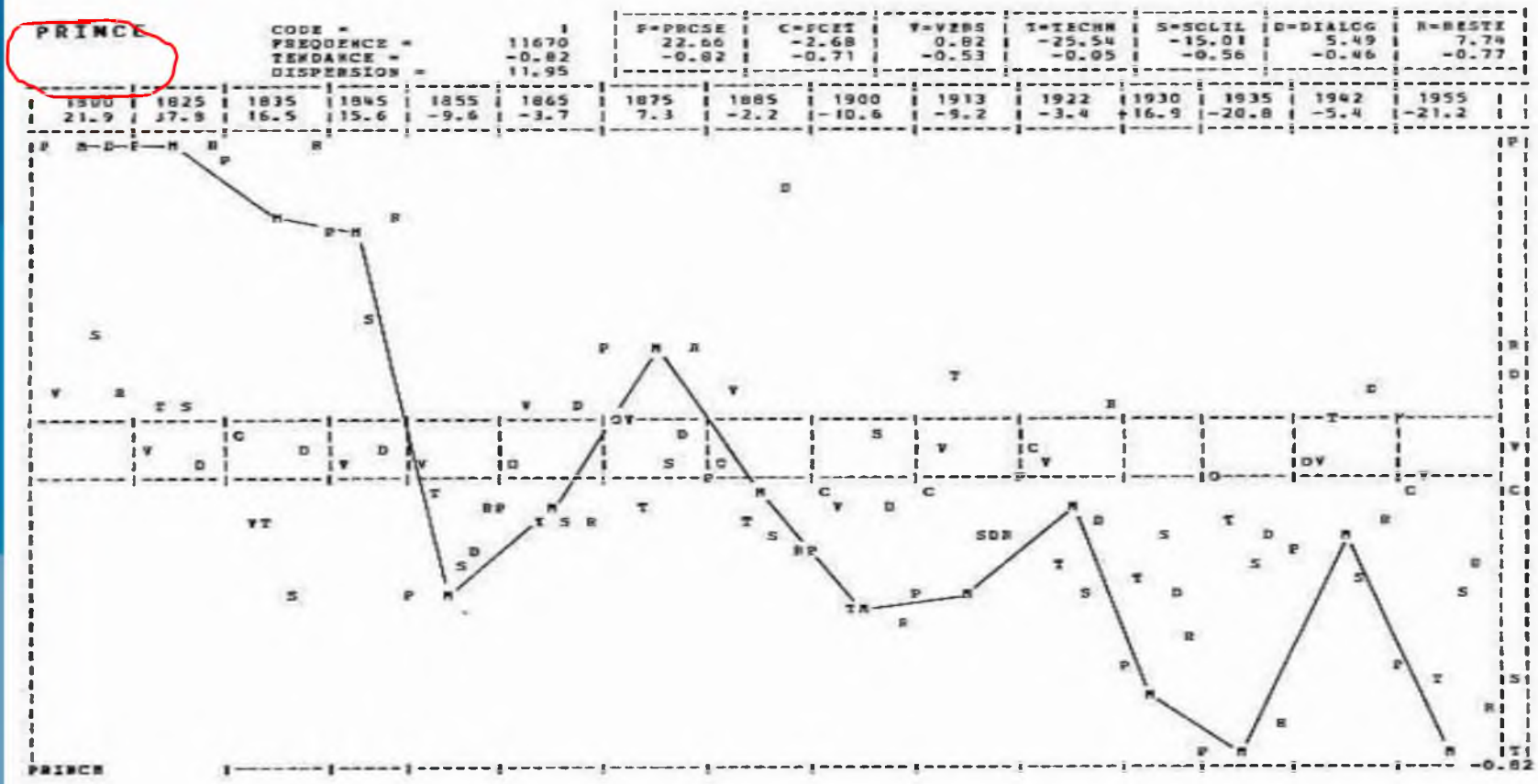
1) Cykle gospodarce traktowane są w tradycji ekonomii autonomicznie, brak jednak ich głębokiego uzasadnienia (pochodna polityki?, astronomii?)
2) Cykliczność w konstrukcji obrazu świata wyrasta z tradycji mistycznych, filozoficznych (np. pitagoreizm) i religijnych (por. biblijny sen o siedmiu krowach)

WZORCE SEKWENCJI: STABILNY TREND



STABILNY TREND W PRAKTYCE

Przykład (prawdopodobnie) pierwszych szeregów leksykalnych w lingwistyce. Leksem *książę* (fr. prince) w tekstach francuskich od 1800 do 1960 r. (frekwencje znormalizowane)



Źródło: E. Brunet (1981), *Le vocabulaire français. De 1789 à nos jours*. Paris, Genève: Slatkine, Champion

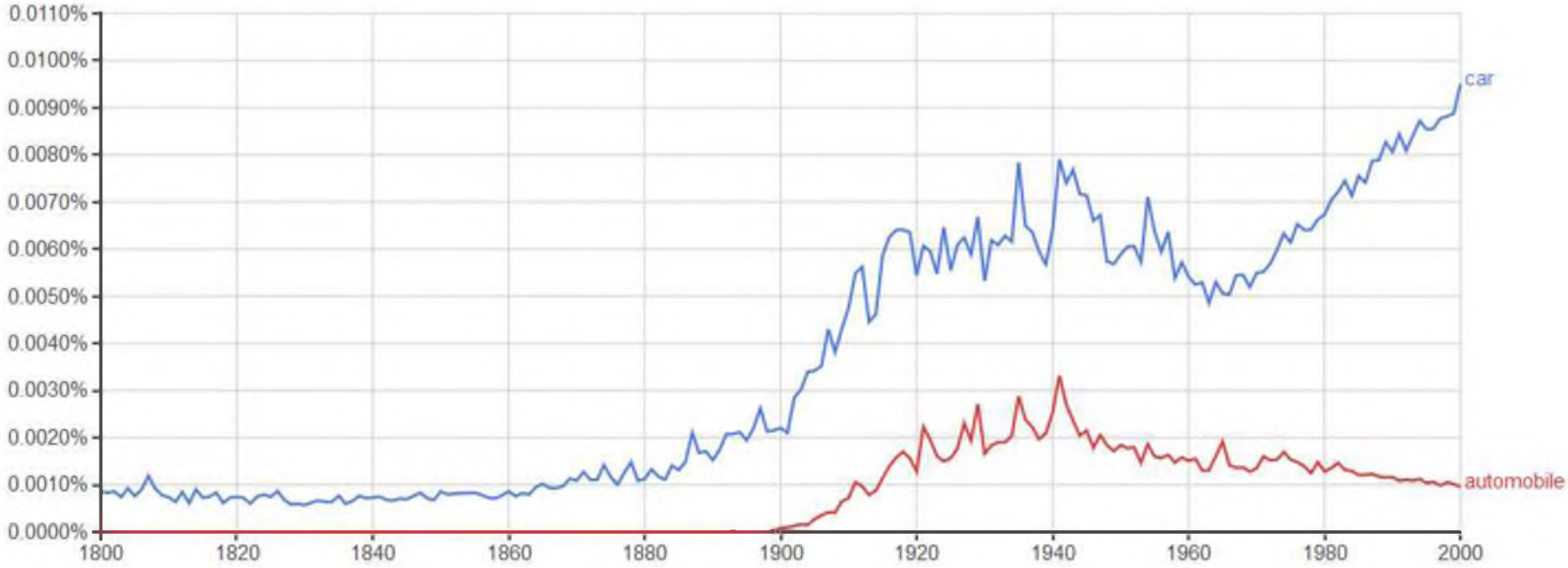
Uniwersytet Wrocławski

STABILNY TREND W GOOGLE BOOKS

Google books Ngram Viewer

Graph these comma-separated phrases: car,automobile ☐ case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 0 Search lots of books



Zastrzeżenia:

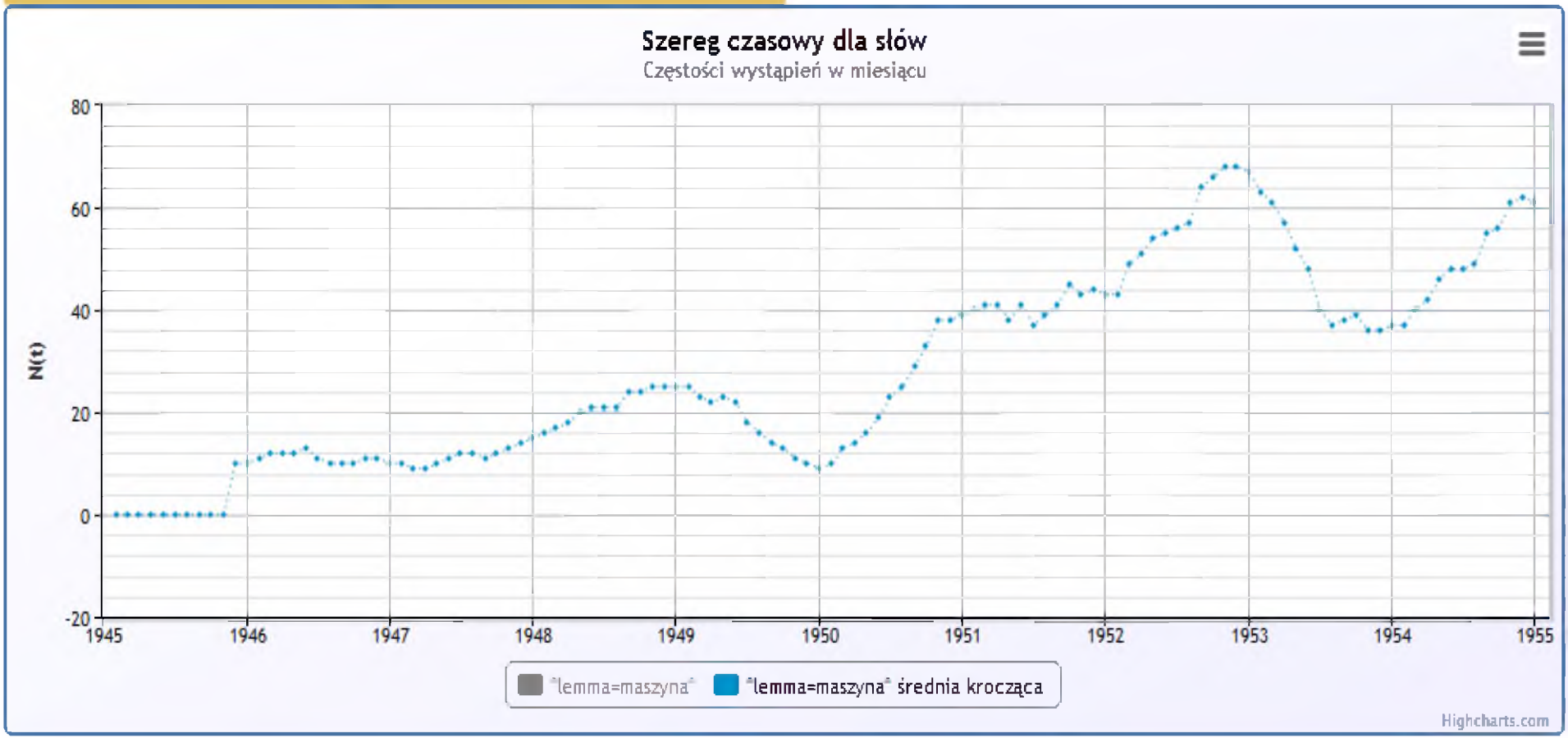
1) Niejasna kwestia kompletności danych (baza?, języki?)

2) Kiedy można mówić o trendzie stabilnym, monotonicznym?

Uniwersytet Wrocławski

Stabilny trend w korpusie ChronoPress

<lemma=maszyna> ✖ Query

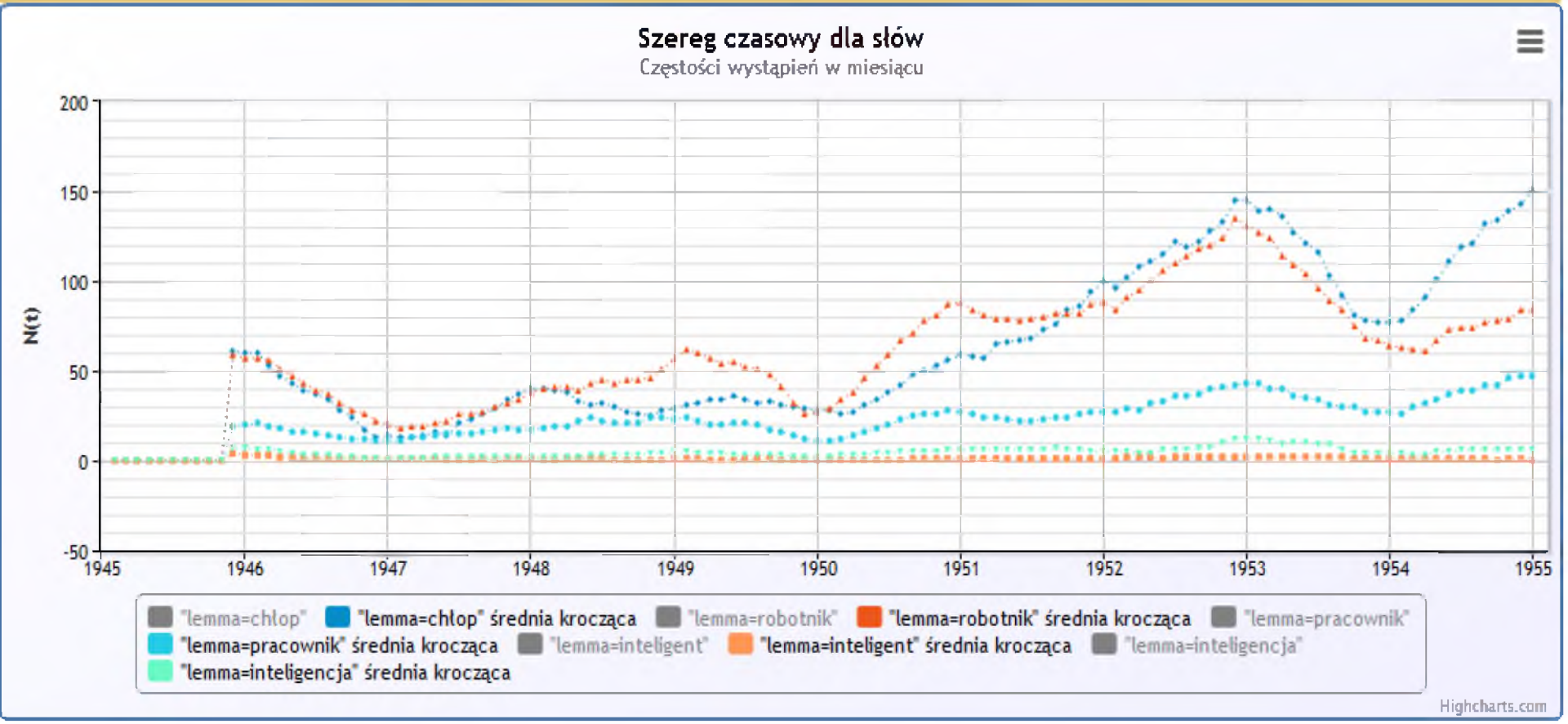


Uniwersytet Wrocławski

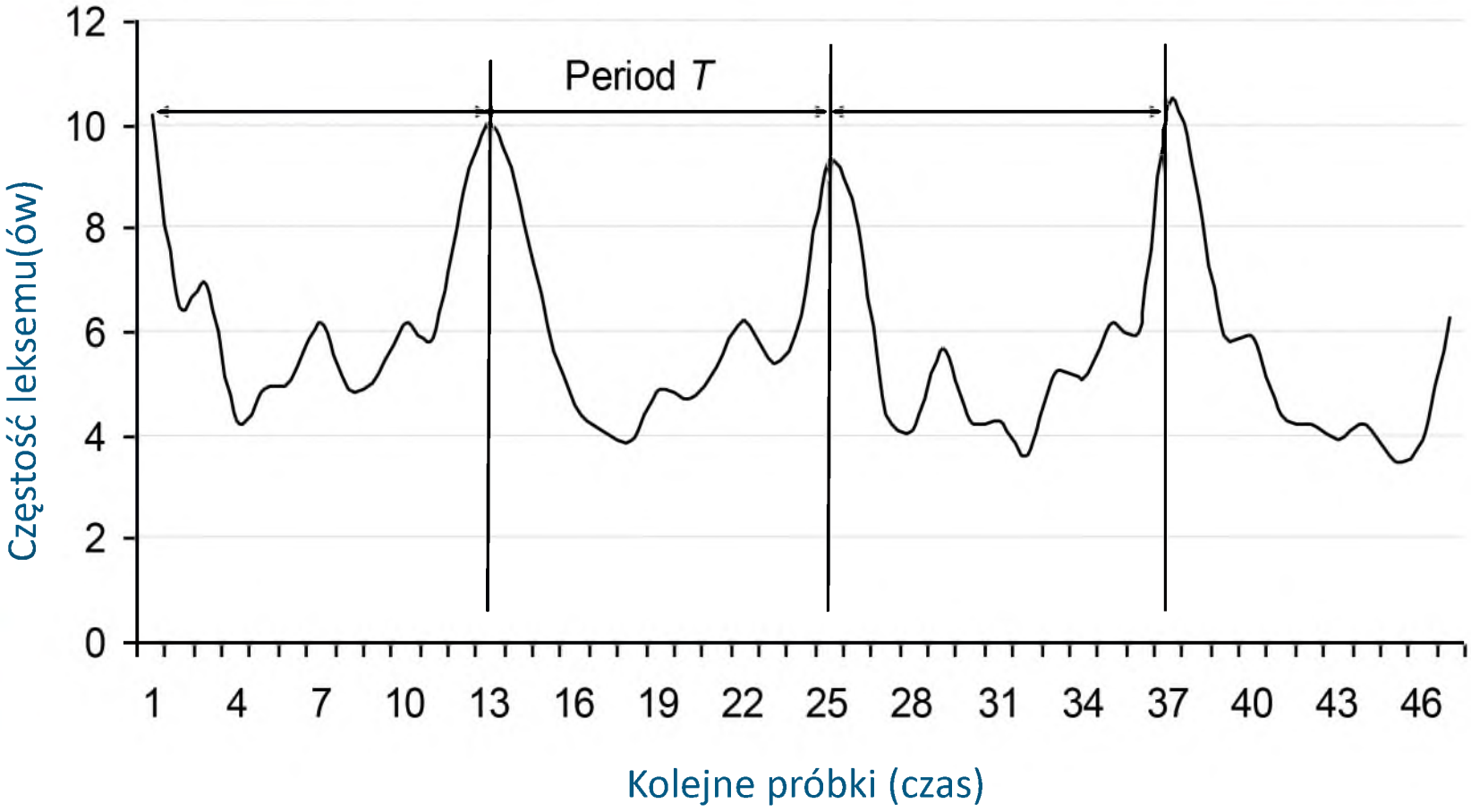
Stabilny trend w korpusie ChronoPress

<lemma=chłop> ✖ <lemma=robotnik> ✖ <lemma=pracownik> ✖ <lemma=inteligent> ✖

<lemma=inteligencja> ✖ Query

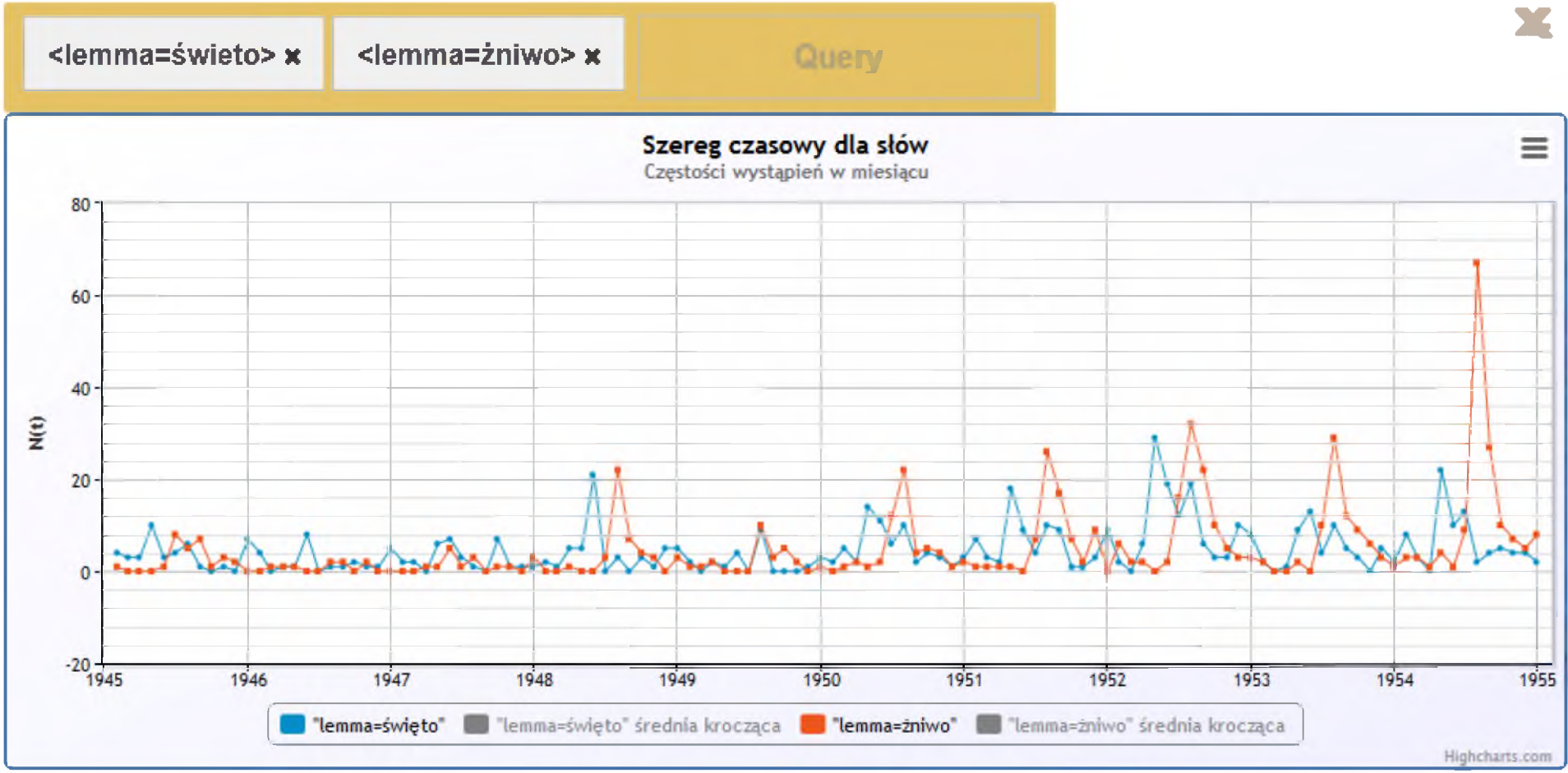


WZORCE SEKWENCJI: SCHEMAT OSCYLACJI



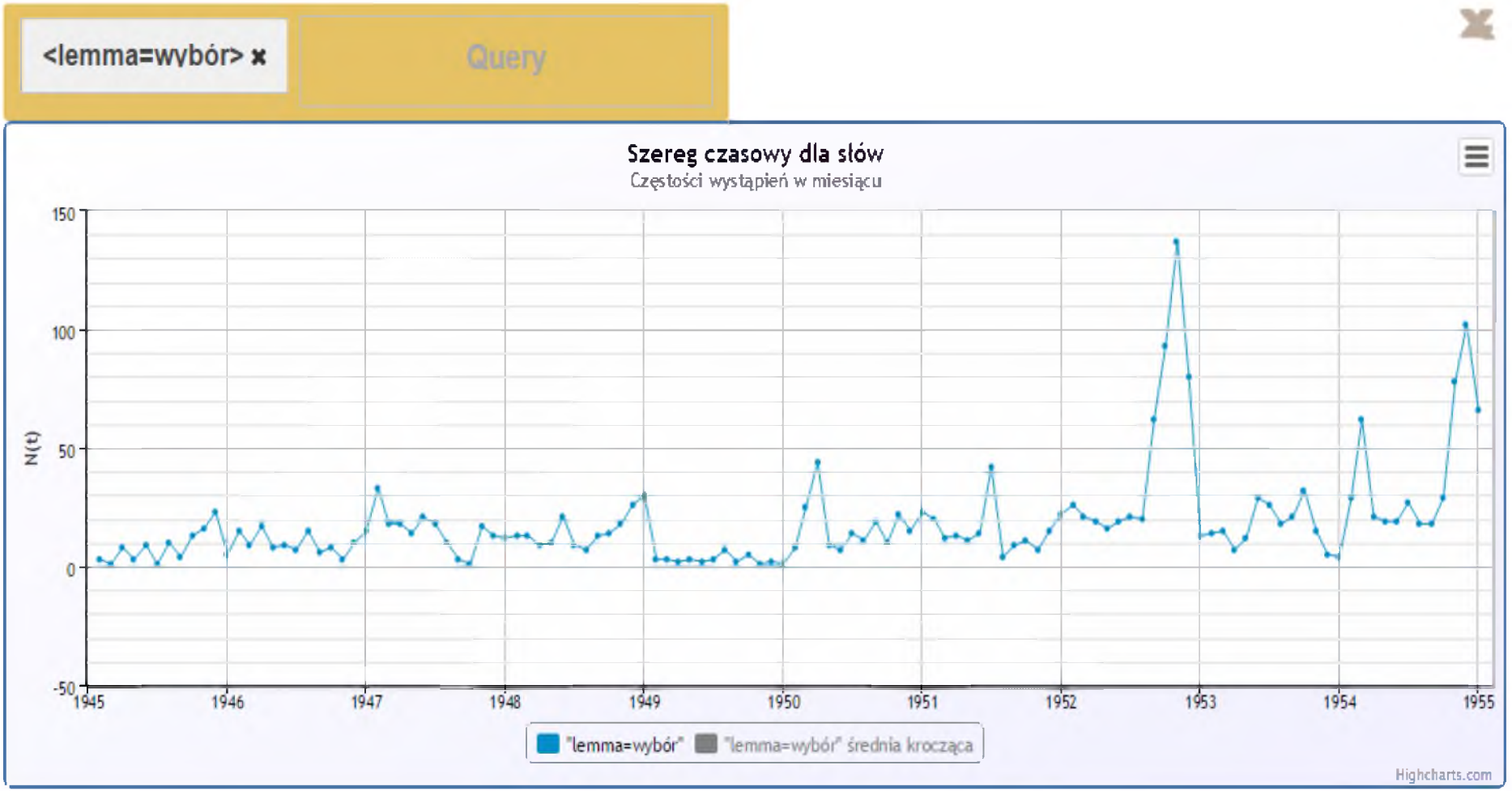
Uniwersytet Wrocławski

PRZYKŁAD OSCYLACJI PERIODYCZNYCH

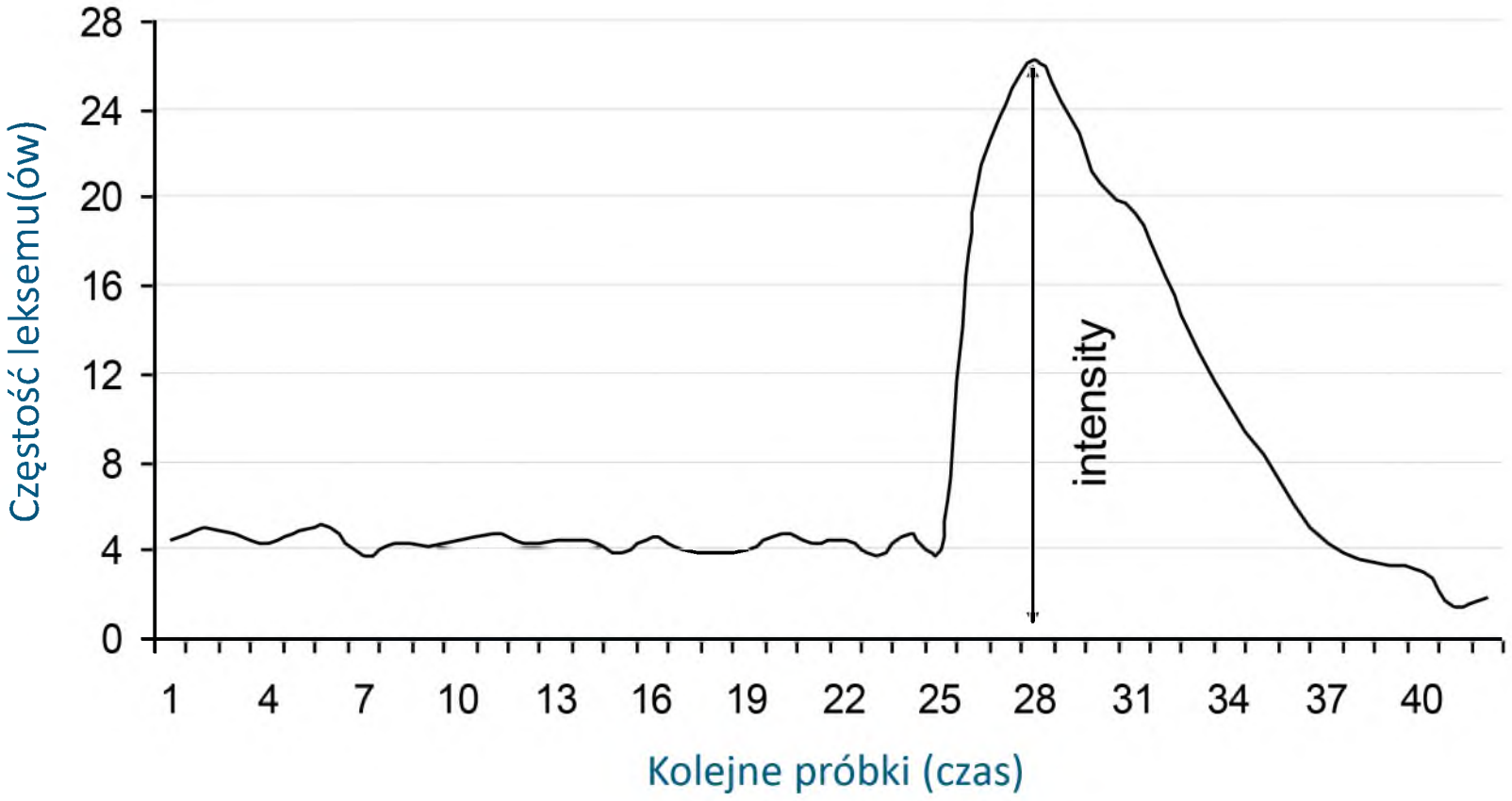


Uniwersytet Wrocławski

PRZYKŁAD OSCYLACJI PERIODYCZNYCH



WZORCE SEKWENCJI: „KATASTROFA”



Uniwersytet Wrocławski

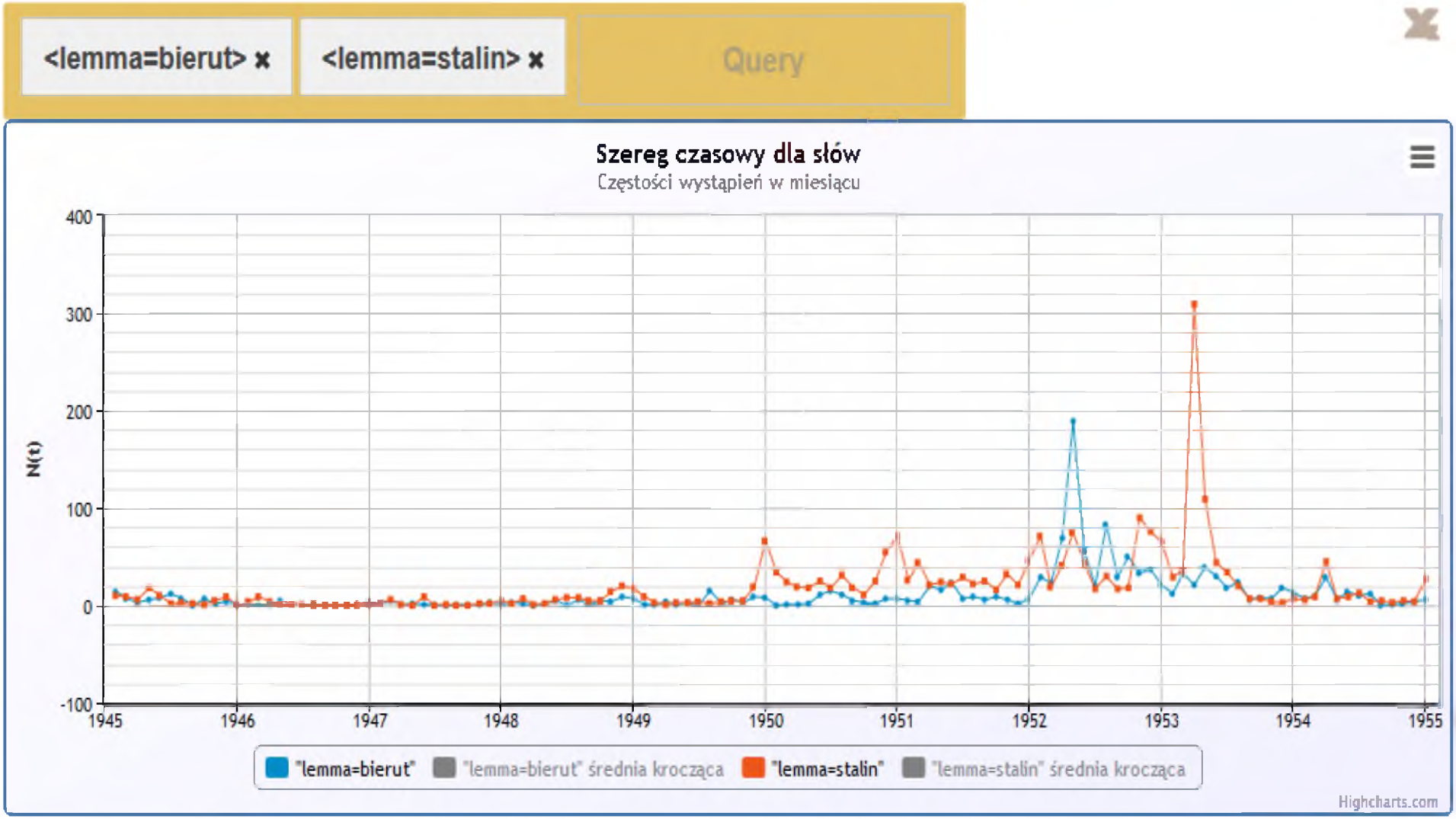
WZORCE SEKWENCJI: „KATASTROFA”

(<lemma=naród> <lemma=polski>)~2 ✖ (<lemma=naród> <lemma=radziecki>)~2 ✖

Query

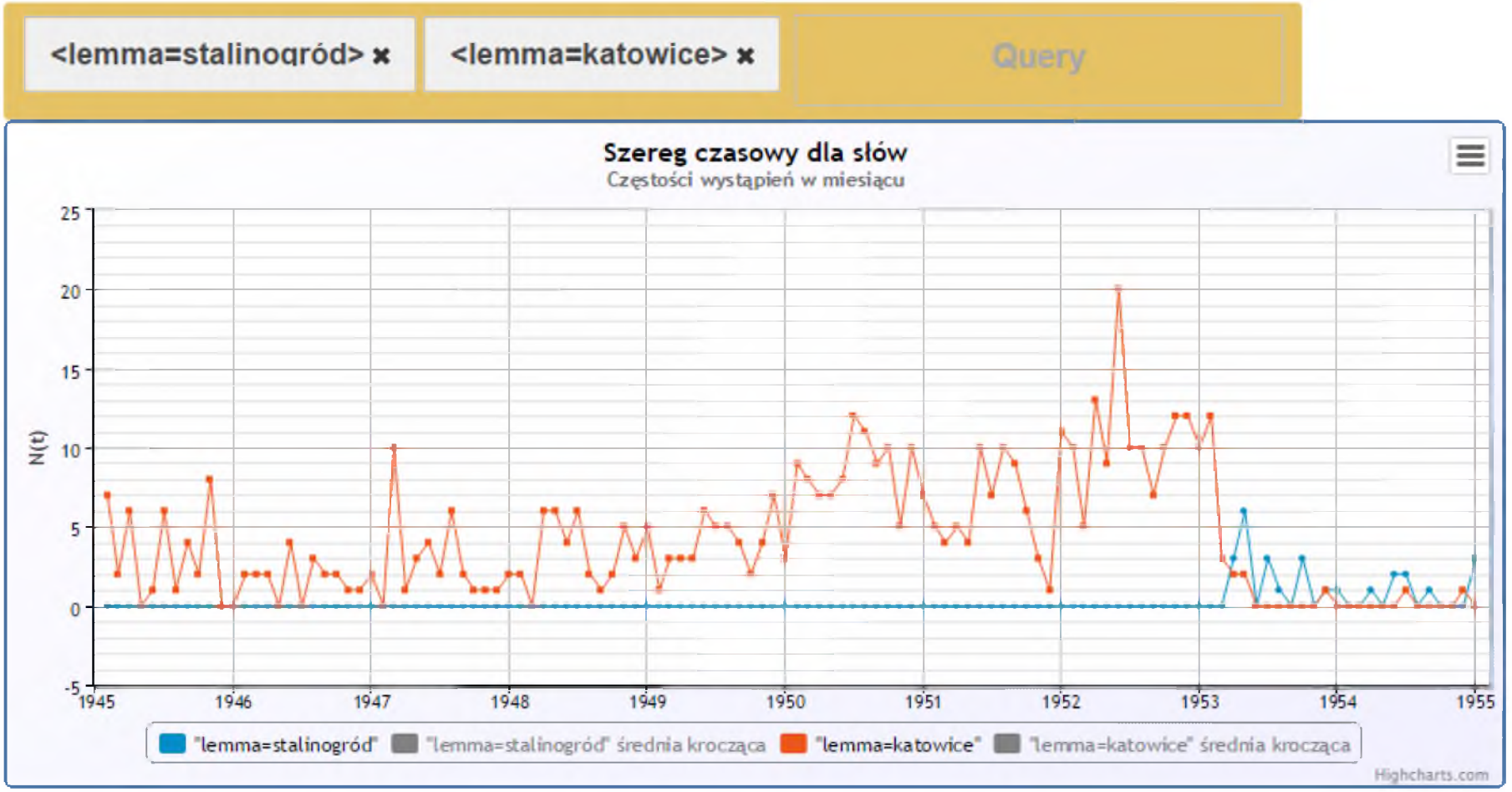
Uniwersytet Wrocławski

WZORCE SEKWENCJI: „KATASTROFA”



Uniwersytet Wrocławski

WZORCE SEKWENCJI: „KATASTROFA”



Uniwersytet Wrocławski

SCHEMAT KATASTROFY (N-GRAM VIEWER)

Google books Ngram Viewer

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

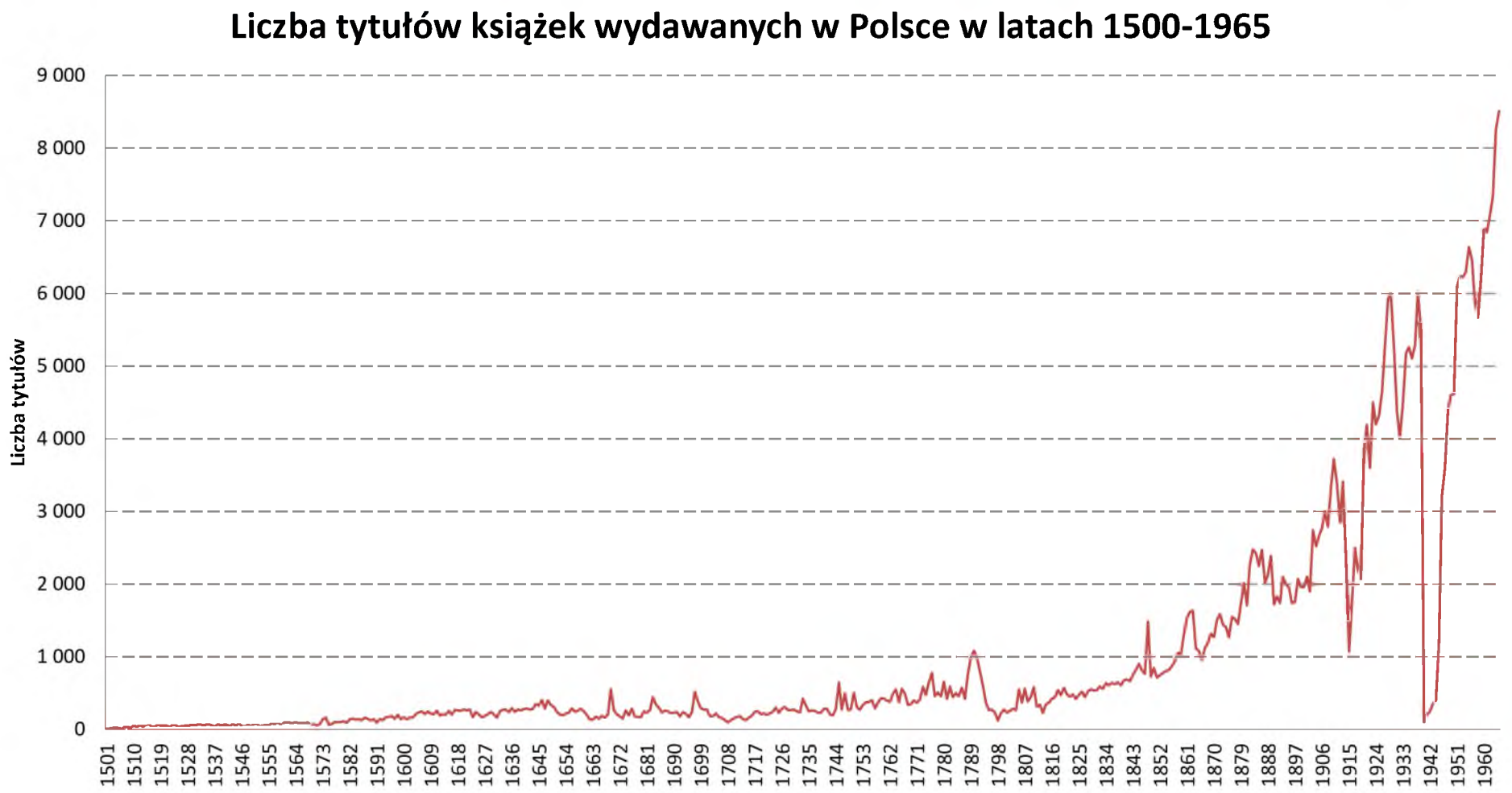
between 1800 and 2008 from the corpus English with smoothing of 0 Search lots of books

PRZYROST PRODUKCJI WYDAWNICZEJ

dr Maria Cecylia Czarnowska (1906–2001)

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.

Inne miary: „Katastrofa” w produkcji wydawniczej



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

SŁOWO MIESIĄCA A BADANIA CHRONOLOGICZNE

Odpowiednikiem kategorii „słowo miesiąca/roku” jest leksykalny wyznacznik „katastrofy”

Katastrofę należy rozumieć tutaj jako zaburzenie stabilnego rytmu zdarzeń, niekoniecznie postrzegane jako czynnik negatywny.

Słowo miesiąca / roku jest z technicznego punktu widzenia jednostką leksykalną o zwiększonej (czyli pozytywnej) frekwencji.

Natomiast katastrofą jest także zniknięcie jednostki z dyskursu, a więc słowo o frekwencji zerowej (przykład Katowic w 1953 r.).

<u>Różnice</u>

Kryterium selekcji „katastrofy” nie jest sama wartość frekwencji leksemu, lecz jej gwałtowna zmiana.

Uniwersytet Wrocławski

ChronoPress: twórcy

Korpus powstał w ramach konsorcjum CLARIN-PL.
http://clarin-pl.eu/

Dane pozyskuje i opracowuje zespół ok. 60 osób (studentów, doktorantów).

Oprogramowanie powstało w różnych grupach skupionych w Clarin-PL, jednak przede wszystkim w zespole dr. Piotra Pęzika na UŁ i w zespole dr. Macieja Piaseckiego na PWr.

Projekt finansowało Ministerstwo Nauki i Szkolnictwa Wyższego.

Uniwersytet Wrocławski

ChronoPress: stan obecny i perspektywy

Okres: 1945-1954

Objętość: ok. 5760 próbek /rok

Stan wypełnienia operacyjnego: 50%

Reprezentatywność: prasa oficjalna

Rozbudowa pionowa (kolejne okresy)

- 1945-1954-1990
- 1939-1944
- 1918-1939

Rozbudowa pozioma (nowe języki):

- czeski
- niemiecki
- inne języki

Rozwój nowych funkcjonalności

Uniwersytet Wrocławski

ChronoPress: zawartość 1950-1954

Lp	Tytuł	Słów na mc	Proc.	Numerów/mc	Słów/nr	Prób/mc	Rok
1	Trybuna Ludu	24000	20%	30	800	96	1 152
2	Trybuna Robotnicza	6000	5%	26	231	24	288
3	Gazeta Robotnicza (L)	6000	5%	26	231	24	288
4	Sztandar Młodych	12000	10%	26	462	48	576
5	Żołnierz Wolności	12000	10%	26	462	48	576
6	Gromada	6000	5%	12	500	24	288
7	Chłopska Droga	6000	5%	4	1 500	24	288
8	Zielony Sztandar	6000	5%	4	1 500	24	288
9	Przekrój	12000	10%	4	3 000	48	576
10	Życie Warszawy (L)	6000	5%	26	231	24	288
11	Tygodnik Powszechny	12000	10%	4	3 000	48	576
12	Przyjaciółka	12000	10%	4	3 000	48	576
13	Kobieta i Życie	0	0%	0	0	0	0
14	Ekspres Wieczorny	0	0%	0	0	0	0
15	Przegląd sportowy	0	0%	0	0	0	0
Suma		120 000	100%	192		480	5 760

ChronoPress: zawartość 1945

Nr	Tytuł	Słów na mc	Proc.	Numerów/mc	Słów/nr	Prób/mc	Rok
1	Głos Ludu	12000	7%	30	400	48	576
2	Robotnik	12000	7%	26	462	48	576
3	Rzeczpospolita	12000	7%	26	462	48	576
4	Trybuna Robotnicza/Śl. (L)	6000	4%	26	231	24	288
5	Dziennik Polski	6000	4%	27	222	24	288
6	Nowe Życie	6000	4%	28	214	24	288
7	Walka Młodych	6000	4%	26	231	24	288
8	Pionier (L)	6000	4%	26	231	24	288
9	Gazeta Lubelska	6000	4%	27	222	24	288
10	Słowo Pomorskie						
11	Kurier Szczeciński						
12	Wolna Łódź						
13	Wiadomości Szczecińskie Pionier Szczeciński						
14	Wiadomości Bydgoskie						
15	Polska Zbrojna	12000	7%	26	462	48	576
16	Zwyciężymy	12000	7%	27	444	48	576
17	Orzeł Biały	6000	4%	28	214	24	288
18	Wolna Polska	6000	4%	29	207	24	288
19	Wolność	12000	7%	26	462	48	576
20	Chłopi	6000	4%	26	231	24	288
21	Chłopska Droga	6000	4%	4	1 500	24	288
22	Wieś						
23	Zielony Sztandar	6000	4%	4	1 500	24	288
24	Przekrój	12000	7%	4	3 000	48	576
25	Życie Warszawy (L)	6000	4%	26	231	24	288
26	Tygodnik Powszechny	12000	7%	4	3 000	48	576
27	Gość Niedzielny						
	(Ekspres Wieczorny)	0	0%	0	0	0	0
	(Przegląd sportowy)	0	0%	0	0	0	0
Suma		168 000	100%	446		672	8 064

Składnia zapytań w aktualnej wersji ChronoPressu

#	Query	Returns text spans containing
1	mamo	A single surface token
2	wiesz co	A sequence of surface tokens
3	<lemma=palić>	All variants of a single lemma token
4	<lemma=mieć> <lemma=szansa>	A sequence of immediately adjacent lemma tokens
5	słuchaj <lemma=ja>	A sequence of surface and lemma tokens
6	tu|tutaj	Variants separated by the pipe operator[6]
7	<lemma=facet>|<lemma=koleś>	Lemma variants
8	bardzo|strasznie dużo	Surface token variants in a sequence
9	(ta kobita)=1	Sequence separated by zero or one unspecified tokens
10	(<lemma=jechać> tam)=2	A lemma and a surface token separated by up to 2 tokens
11	(<lemma=jechać> tam)~2	As above except that the tokens may occur in any order
12	(<lemma=dać> do zrozumienia)=2	3 obligatory tokens separated by up to 2 unspecified ones
13	<lemma=prosić>|!proszę	Any form of *prosić* except for *proszę*[7]
14	t.* bab.*	Tokens matching regular expressions
15	szykow.+|przygotow.+	Variants with regular expressions
16	<lemma=p.+biec>	Lemmas matching a regular expression
17	<tag=subst:pl:.*>	Any plural noun
18	<tag=subst:.+:inst:.+>	Any noun in the instrumental case
19	<lemma=zdać pos=verb:sg:.*>	Singular forms of the verb "zdać"
20	<tag=adj:.+> <lemma=temat>	Sequences of adjectives preceding the noun "temat"
21	(<lemma=słuchać> <tag=.*:gen:.*>)=1	Lemma followed by any genitives with a slop factor

źródło: Pęzik Piotr (2014), Spokes – a search and exploration service for conversational corpus data. CLARIN 2014 Selected Papers. Linköping Electronic Conference Proceedings # 116, p.104

Uniwersytet Wrocławski

Wykonawcy: CLARIN-PL

http://clarin-pl.eu/

Prof. Adam Pawłowski, Uniwersytet Wrocławski

Uniwersytet Wrocławski

Obecna strona ChronoPress

http://clarin.pelcra.pl/chronopress/

Prof. Adam Pawłowski, Uniwersytet Wrocławski

Uniwersytet Wrocławski

DZIĘKUJĘ ZA UWAGĘ

Prof. Adam Pawłowski, Uniwersytet Wrocławski