

# Korpusy - zastosowania

Beata Wójtowicz

**Prezentacja chroniona prawem autorskim.**

---

## 3 podejścia w badaniach w językoznawstwie korpusowym:

1. ilustrowane przez korpus (ang. *corpus illustrated*),
  2. oparte na korpusie (ang. *corpus based*),
  3. sterowane korpusem (ang. *corpus driven*).
-

---

# 1. ilustrowane przez korpus

W argumentacji posługujemy się przykładami z korpusu zamiast tworzonych ad hoc. Korpus pełni rolę taką samą jak informator – rodzimy użytkownik języka.

---

## 2. oparte na korpusie *corpus based*

Badania, które mają na celu zweryfikowanie postawionej wcześniej hipotezy, do sprawdzenia teorii oraz opisów, które były sformułowane przed dostępem do wielkich korpusów.

Nie poszukujemy pojedynczej ilustracji postawionej tezy (jak w 1.), ale przeszukujemy cały korpus po to, by tę tezę z jednej strony potwierdzić bądź sfalsyfikować, z drugiej, by stwierdzić czy nie jest ona pozbawiona luk, a po trzecie, by określić zasięg zjawiska, czy jest typowe czy marginalne.

---

### 3. sterowane korpusem *corpus driven*

Przed przystąpieniem do badań nie stawiamy żadnych hipotez.

Istotą badania jest taksonomiczne budowanie gramatyki na podstawie danych empirycznych pozyskiwanych z korpusu. Od takiego podejścia oczekuje się opisu neutralnego pod względem teoretycznym. Wnioski wyciągnięte z analizy muszą odzwierciedlać tylko i wyłącznie poświadczenia korpusowe.

---

---

U wielu badaczy badania typu 2. i 3. nie są wyróżniane i nazywa się je badaniami opartymi na korpusie (ang. *corpus based*).

R. L. Górski. 2012. Zastosowanie korpusów w badaniu gramtyki. W: *Narodowy Korpus Języka Polskiego*. PWN.

---

# Badanie korpusowe w praktyce

Po części weryfikacja przyjętych hipotez, po części badanie całkowicie indukcyjne.

Analiza raczej ilościowa niż tylko jakościowa, bardziej poszukiwanie tendencji niż reguł.

Rozpatrywanie równocześnie formy i funkcji np. bytów gramatycznych.

---

# Metoda a cel badań

Wybór metody związany jest z celem badania.

Jeśli zamierzamy sprawdzić określoną teorię, to używamy metody *corpus-based*.

Badanie: dystrybucja końcówek w mianowniku l.mn. w wyrazie *profesor*.

Corpus-based: szukamy form *profesorowie* i *profesorzy*, porównujemy wyniki frekwencji + analizujemy poświadczenia.

Corpus-driven: nie szukamy konkretnych form, tylko poświadczeń wyrazu w mianowniku l.mn.

---



---

# Metoda a cel badań

W badaniach semantycznych, np. analizując stereotypy.

Stereotyp Polaka

metodą *corpus-based*: wyszukujemy znane nam cechy, jak *złodziej*, *pijak*;

metodą *corpus-driven* układamy cechy wedle poświadczeń korpusowych.

---

---

# *Corpus-driven*

## **Profile kolokacyjne** (*collocation profiles*)

Tworzenie wizerunku leksemu za pomocą grupowania wyrazów współwystępujących.

## **Pattern grammar**

Badanie wzajemnego stosunku danej jednostki leksykalnej do wzorca gramatycznego, którego część stanowi.

*be spent* – jednostkę poprzedza rzeczownik odnoszący się do jednostki czasu (miesiąc, tydzień, rok), natomiast po niej występuje czasownik zakończony na *-ing*.

---

# Podstawowe zasady

Dobór korpusu do celu badań.

Wybór właściwej metody badawczej.

Poznanie teorii językowej, na podstawie której powstała anotacja.

Przewidywanie błędów w anotacji.

Prawidłowe konstruowanie zapytań –  
znajomość anotacji, języka zapytań ->

NKJP- wyszukiwarka PELCRA nie robi rozróżnienia na małe i wielkie litery – zapytanie *brat* zwraca wszystkie poświadczenia

wyszukiwarka Poliqarp ma takie rozróżnienie, w celu pozyskania wszystkich poświadczeń należy użyć zapytania *brat|Brat*

---

# Monitorowanie zmian w języku

<https://collins.co.uk/pages/elt-cobuild-reference-the-collins-corpus>

The Collins Corpus is so large, we can look at lots of examples of how people really use the words. The data tells us how words are used, what they mean, which words are used together, and how often words are used. This information on frequency helps us decide which words to include in the COBUILD dictionaries.

---

# Monitorowanie zmian w języku

Phrase 'in the cloud', how its meaning has changed over time:

1990-1999: 100% of all examples of 'in the cloud' in the written Subcorpus relate either to meteorology or clouds of dust, smoke or alike.

2005-2009: More than 60% of the entries found in the Corpus relate to computing.

- The compound 'cloud computing' didn't even exist in the 1990-1999 Corpus, but there are more than 500 hits in the 2005-2009 Corpus.

This shows that the word 'cloud' was mainly used in its meteorological sense before 2005. In recent years, 'cloud', in the sense of web-based storage for files, has become frequent along with new collocations ('cloud computing') and new phrases ('in the cloud').

---

# Korpus w leksykografii

- Ustalenie makrostruktury,
  - walka z różnego rodzaju niekonsekwencjami,
  - odnalezienie form derywacyjnych i sprowadzenie ich do wspólnego rdzenia,
  - opracowywanie treści artykułu hasłowego (konkordancje),
  - znalezienie typowych kolokacji, idiomów i przysłów,
  - znalezienie/ustalenie najbardziej typowych, przykładów użycia.
-

---

Jakie znaczenia posiada dany wyraz?

Jak często używany jest dany wyraz?

Jak wygląda jego użycie w różnych rejestrach języka?

W towarzystwie jakich innych wyrazów najczęściej występuje dane słowo?

Jak dystrybuowane są poszczególne znaczenia i użycia wyrazu? jak częste są poszczególne znaczenia czy funkcje gramatyczne danego wyrazu? czy te znaczenia są odnotowane we wszystkich rejestrach języka?

Jak różnią się użycia synonimów? czy łączą się w różne kolokacje? czy są używane tak samo w różnych rejestrach języka?

---

# Siatka haseł a lista frekwencyjna

|             | Johnson | Perrott | Stopa i Garlicki | TUKI | Abdulla |
|-------------|---------|---------|------------------|------|---------|
| sema        | T       | T       | T                | T    | T       |
| la          | T       | T       |                  | T    | T       |
| nchi        | T       | T       | T                | T    | T       |
| mtu         | T       | T       | T                | T    | T       |
| ingine      | T       | T       |                  | T    | T       |
| fanya       | T       | T       | T                | T    | T       |
| toa         | T       | T       | T                | T    | T       |
| kwenye      | T       | T       |                  |      | T       |
| rais        |         |         |                  | T    | T       |
| kiongozi    | T       | T       | T                | T    | T       |
| waziri      | T       | T       | T                | T    | T       |
| mkoa        |         | T       |                  | T    | T       |
| polisi      | T       | T       |                  | T    | T       |
| endelea     | T       | T       |                  | T    | T       |
| taifa       | T       | T       |                  | T    | T       |
| kutokana na |         |         |                  |      | T       |
| anza        | T       | T       | T                | T    | T       |
| mwananchi   | T       | T       |                  |      | T       |
| kutoka      |         |         |                  |      | T       |
| mahakama    | T       |         |                  | T    | T       |
| zaidi       | T       | T       | T                | T    | T       |
| baadhi      | T       | T       |                  | T    | T       |
| uchaguzi    | T       | T       |                  | T    | T       |



---

# TUKI - kolokacje

Słownik: cheza mpira; ~ na mtoto

Korpus: cheza ngoma, ~ mchezo, ~ dansi

---

---

# Inny Słownik Języka Polskiego

Definicje kontekstowe oparte na analizie łączliwości gramatycznej i leksykalnej wyrazów.

Autentyczne przykłady.

Warianty haseł, podhaseł i form fleksyjnych uporządkowane są według ich frekwencji.

---

# Definicje i przykłady

## Ładować

Mówimy, że ktoś **ładuje** w coś pieniądze, jeśli inwestuje je w to. Słowo potoczne. *Harował przez całe życie, a co zarobił, ładował w maszyny... Każdy grosz ładowała w unowocześnienie swojego sklepu.*

## Taniocha

Jeśli gdzieś jest **taniocha**, to wszystko jest tam bardzo tanie. Słowo potoczne. *Wyjeżdżamy na Kaukaz, gdzie podobno taniocha.* > Także o czymś bardzo tanim. *Szalupa kosztuje sześć kawałków, taniocha.*

---

# Warianty

Warianty *jazz* (334 wystąpienia) i *dżez* (9 wystąpień)

ISJP – *jazz* hasło główne, *dżez* opisano jako hasło odsyłaczowe z kwalifikatorem 'rzadziej'.

Inne słowniki – odwrotnie – pierwszeństwo ma spolszczona forma *dżez* (np. SWJP, Dunaj 2000).

Warianty niewystępujące w korpusie ISJP pomija, np. spolszczoną formę *rekwiem* – wymieniona jest tylko zaświadczona forma *requiem*.

---

---

# Gramatyka

Dane **statystyczne** (frekwencja danej konstrukcji czy formy fleksyjnej).

Związki gramatyki ze zjawiskami **pozagramatycznymi** – leksyką, pragmatyką, dyskursem, stylistyką

**Konkurencja** blisko- lub jednoznacznych form i konstrukcji (*uczniów – uczni; mówić do kogoś – mówić komuś*)

Badanie różnicowania **stylistycznego** języka.

---

---

Punktem wyjścia najczęściej jest poszukiwanie konkretnego leksemu.

Wykorzystanie znakowania morfosyntaktycznego.

Badania:

Zdania podrzędne → [base = 'że|iż|żeby']

Wariantywność form fleksyjnych - formy wariantywne  
I.mn. rzecz. męskoosobowych zakończonych na -er  
lub -or (reżyserzy|reżyserowie)

[base = '.\*er|.\*or' & orth = '.\*owie|.\*rzy']

---

# Składnia

Dystrybucja dwu realizacji argumentu czasownika  
*opowiadać: opowiada o tym, że <> opowiada, że*

Możliwa długość ciągu złożonego z samych  
rzeczowników lub przymiotników w dopełniaczu –  
jak bardzo można rozbudować frazę nominalną?

Dystrybucja dwóch postaci czasu przyszłego  
złożonego: *będę pisała <> będę pisać*

---

# NKJP w leksykografii

***Inny słownik języka polskiego***, red. M.Bańko. Warszawa 2000.

***Wielki słownik angielsko-polski PWN***. Warszawa 2004/2005.

***Słownik dobrego stylu, czyli wyrazy, które się lubią***, red. M. Bańko. Warszawa 2006.

M. Czeszewski, K. Foremniak, ***Ludzie i miejsca w języku. Słownik frazeologizmów eponimicznych***, Warszawa 2011.

***Wielki słownik języka polskiego***, red. P.Żmigrodzki ([wsjp.pl](http://wsjp.pl))

---



## Lingwistyczne

Na przykład:

- kontrastywna analiza wyrażania emocji w języku (Dziwirek i Lewandowska-Tomaszczyk, 2010),
- prace nad różnymi zjawiskami gramatycznymi:
  - stroną bierną (Górski, 2008, 2011),
  - konstrukcjami bezokolicznikowymi (Gębka-Wolak, 2011),
  - itp.

Poradnia językowa PWN, np.:

- P: *Zmotywować, zmotywowany, zdemotywować, zdemotywowany.*  
Chciałabym zapytać, czy powyższe formy są poprawne...
- O: Wymienione wyrazy nie są źle zbudowane... Jeśli w słownikach ich nie ma, tym gorzej dla słowników. Narodowy Korpus Języka Polskiego ma ponad 500 przykładów dla *zmotywować*...

## Translatoryczne

Pęzik 2011:

- ocena poprawności tłumaczenia, np.:
  - O: Why *political orthodoxy* must not silence *scientific argument*.
  - T: Dlaczego *ortodoksja polityczna* nie może uciszyć *kłótni naukowej*.
- czy *ortodoksja polityczna* i *kłótnia naukowa* to równie poprawne kolokacje w języku polskim, jak ich odpowiedniki w języku angielskim?
- analiza kolokacji w NKJP pokazuje, że:
  - *ortodoksja polityczna* jest poprawnym tłumaczeniem,
  - ale zamiast *kłótni naukowej* należałoby raczej użyć wyrażenia *debata naukowa*, *spór naukowy* lub *dyskusja naukowa*.

---

Wiele innych:

- podprojekt *Słowa dnia* (Łaziński i Andrzejczuk, 2011) – **monitorowanie częstości słów w prasie** (śledzenie aktualnie poruszanych tematów, znajdowanie nowych słów w języku),
- **pomoc w nauczaniu języka polskiego** (także jako obcego) – NKJP wykorzystywane m.in. na Uniwersytecie Warszawskim i Uniwersytecie Jagiellońskim,
- **pomoc w nauczaniu inżynierii lingwistycznej** (m.in. zajęcia na MIM UW),
- **rozwiązywanie krzyżówek** (*warstwa lodu tworząca się na powierzchni ziemi wskutek zamarzania mżawki lub deszczu; [base="..ło..d." ]*),
- i in.

Najważniejsze w IPI PAN: **tworzenie modeli językowych**  
**i trenowanie narzędzi przetwarzania języka.**

---

---

# Słowa dnia

Słowa dnia i tygodnia to te, których w ostatnim dniu lub tygodniu użyto w gazetach znacznie częściej, niż by się można spodziewać na podstawie frekwencji z okresu porównawczego, mierzonego w miesiącach lub latach. Najczęstsze nazwiska, nazwy miejsc, firm, organizacji, a także rzeczowniki pospolite, czasowniki i przymiotniki, mówią o tym, kto był bohaterem dnia, czym zajmowali się politycy i urzędnicy, na kim i na czym skoncentrowało się zainteresowanie dziennikarzy i czytelników.

Słowa [wczorajszego dnia](#)

---

---

## *Przysłowia mądrością narodu..., czyli jak przy użyciu korpusu dowiedzieć się czegoś o mentalności użytkowników języka.*

Anna Dukaczewska, 2009; dane z witryny korpusy.net

W badaniu za punkt wyjścia posłużyła konstrukcja (dalej zwana też ramą zdaniową): KTO ... TEN...

„kto pierwszy, ten lepszy”

„kto płaci, ten wymaga,”

Jeśli polegać na intuicji - można odnieść wrażenie, że wyrażenia tego typu pojawiają się w języku polskim dość często. Jednak by znaleźć empiryczne potwierdzenie tego sądu, warto skorzystać z korpusu.

Szczególnie ciekawe wydaje się sprawdzenie, ile z „uzupełnień” ramy KTO...TEN... jest rzeczywiście przysłowiami (skonwencjonalizowanymi/utartymi w danej społeczności sformułowaniami wyrażającymi prawdy ogólne), a ile z nich to oparte na znanej ramie wytwory własne rodzimych użytkowników języka.

---

# Wyniki analizy korpusu IPI PAN

## I. Przysłowia/maksymy

*kto pierwszy ten lepszy* – 41 razy na 114 sekwencji

*kto płaci ten wymaga* – 9 wyników

*kto korzysta ten płaci* – 4 wyniki

II. Często powtarzane powiedzenia, stare przysłowia czy ludowe porzekadła. W analizowanym zbiorze sekwencji pojawiają się one jednorazowo.

*kto śpiewa, ten się dwa razy modli*

*kto śpi, ten nie grzeszy*

*kto milczy, ten kłamie*

*kto hoduje, ten poluje*

Każda z wymienionych wyżej maksym służy uregulowaniu jakiejś innej dziedziny życia, ma pomóc w odróżnieniu dobra od zła, czy też narzucić odpowiednie postępowanie.

## III. Radykalne poglądy.

*Kto pali, ten kradnie i może też zamordować*

*Kto prawosławny, ten Rosjanin*

Czy możemy uznać, że wyrażają one poglądy charakterystyczne dla współczesnego Polaka?

•Warto posłużyć się szerokim kontekstem i metadanymi (może odnoszą się do dawniejszych czasów?)

---

# Inne zastosowania

A corpus approach to cultural keywords: a critical corpus-based analysis of ideology in the Blair years (1998-2007) through print news reporting.

- What are the key-words for the years 1998–2007, as evidenced in the British press and can they be identified as cultural keywords?
- Have they developed meanings specific to this period and have these meanings evolved within the period?

Jeffries (2003) - the meaning of water, in the context of the Yorkshire water crisis of 1995.

Jeffries (2007) - the way in which the female body was constructed by women's magazines in 2004.

---

---

# British English vs. American English

The purely linguistic differences such as spelling (colour/color) or morphology (got/gotten);

Leech and Fallon (1992) checked up on the senses in which words were being used -> grouped the differences which were statistically significant into fifteen broad categories. The frequencies of concepts in these categories revealed differences between the two countries which were primarily of cultural, not linguistic difference.

- travel words were more frequent in American English than British English,
- words in the domains of crime and the military were also more common in the American data, as was "violent crime" in the crime category, perhaps suggestive of the American "gun culture".

In general, the findings seemed to suggest a picture of American culture at the time of the two corpora (1961) that was more macho and dynamic than British culture.

---



---

# What were the cultural keywords when you were born?

Google Books ngram corpus

1940-1944: nazi hitler aircraft planes grid tanks radio germans vitamin propaganda aluminum voltage automobile orchestra pilot gear altitude civilian coil polish roosevelt poland rubber outstanding vocational

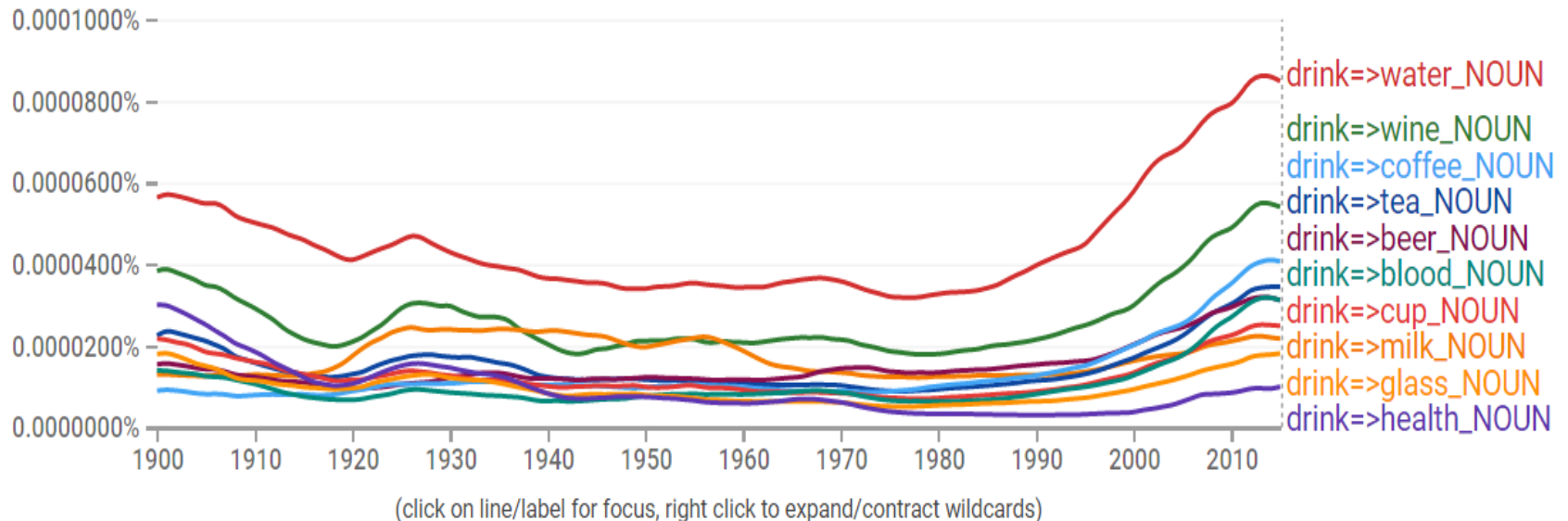
1960-1964: communist communists communism ussr electrons atomic soviet electron radiation equations scattering amino amplitude ions nuclear frequencies equation pakistan coefficients approximation momentum

2000-2004: web global gender options software uk user strategies phone networks

---

# Google Books ngram viewer

Dependencies can be combined with wildcards. For example, consider the query *drink=>\*\_NOUN* below:



---

# Seks w 2005 miał kryzys, ale poprawia wyniki..., czyli jakich słów używa polska prasa

Karolina Głowacka, TOK FM 2012-05-28

---

---

# Keywords project

Narodowy Fotokorpus Języka Polskiego

---

Szanowni Państwo,

niniejsza strona przedstawia największy zbiór leksykalny polszczyzny XX w. wraz z poświadczeniem cytatomym. Jego naczelną dystyngtywną zasadą jest każdorazowe udokumentowanie ekscerptu w postaci fotodokumentacyjnej, tzn. w takiej, w jakiej ukazał się on w druku nim wszedł w nasze e-posiadanie.

Obserwacja leksykalna objęła okres 1901–2000, zaś za kryterium wyboru przyjęto brak jednostki w pierwszym słowniku ortograficznym XX wieku – Słowniku ortograficznym języka polskiego Władysława Kokowskiego z roku 1903.

Za sprawą przyjętych założeń baza ta poważnie uzupełnia obecną wiedzę o zasobach leksykalnych polszczyzny, ponieważ otrzymany zbiór jednostek obejmuje pozycje w dużej mierze w ogóle nieznane polskiej leksykografii lub takie, dla których wysuwano późniejsze hipotezy chronologizacyjne.

Zapraszamy do zapoznania się,





## przedkeynesowski

Znawca teorii **przedkeynesowskiej** mógłby jedynie być oszołomiony tą mnogością konfrontowanych przez siebie wyjaśnień, z których każde było oczywiście wyjaśnieniem konkurującym z innymi, każde oczywiście nie powiązane z innymi i niezależne od innych.

 Gillman, Joseph M. 1969. Kryzys amerykańskiej prosperity, tłum. Z. Chodkiewicz, Warszawa : KiW

## podkonsumpcjonista

Marks w swojej krytyce Saya zadal sobie trud wykazania, dlaczego stosunek między produkcją i konsumpcją nie może być traktowany ani w sposób jednostronny, jak to czynili **podkonsumpcjoniści**, ani w sposób, w jaki traktuje go Say. Gospodarka kapitalistyczna, dowodził Marks, jest gospodarką opartą na wyzysku. Zatrudniając robotników najemnych, kapitalista

 Gillman, Joseph M. 1969. Kryzys amerykańskiej prosperity, tłum. Z. Chodkiewicz, Warszawa : KiW

## właściciel-inwestor

Tę niepewność przewidywań przedsiębiorców, dowodził Keynes, powiększa fakt, że w warunkach korporacyjnego systemu przedsiębiorstw kapitalistycznych własność, czyli funkcja inwestowania, w coraz większym stopniu oddziela się od funkcji przedsiębiorcy. Ci **właściciele-inwestorzy** są przesiąknęci psychologią rynku walorów, która jest wypadkową gwałtownych wahań między spekulacją nadmiernie optymistyczną a spekulacją nadmiernie pesymistyczną. W rezultacie, nastroje te rozlewają się na psychologię przedsiębiorców, a zatem i na przewidywane przez nich zyski.

 Gillman, Joseph M. 1969. Kryzys amerykańskiej prosperity, tłum. Z. Chodkiewicz, Warszawa : KiW

## niewydatkowanie

Oszczędności indywidualne mogą wzbogacić jednostkę. Lecz jeśli nie zostaną zainwestowane, wówczas zubażają naród. Akt oszczędzania jest aktem **niewydatkowania**. **Niewydatkowanie** obniża ceny dóbr konsumpcyjnych, a prawdopodobnie także krańcową wydajność kapitału. Popyt na nowe inwestycje wówczas maleje.

 Gillman, Joseph M. 1969. Kryzys amerykańskiej prosperity, tłum. Z. Chodkiewicz, Warszawa : KiW

## niefinansowy

spokaja, i to z nadwyżką, zarówno krótkoterminowe, jak i długoterminowe zapotrzebowanie na kapitał przy najniższej stopie procentowej w całym świecie kapitalistycznym. Miara tej obfitości może być fakt, że w 1962 r. procenty płacone przez 100 największych **niefinansowych** korporacji w USA wyniosły nie więcej niż 0,7% ich globalnych przychodów, a w 1963 r. — 0,5%.

 Gillman, Joseph M. 1969. Kryzys amerykańskiej prosperity, tłum. Z. Chodkiewicz, Warszawa : KiW

## menedżerstwo

szego Dorocznego Posiedzenia Rady Gubernatorów Międzynarodowej Korporacji Finansowej, oświadczył: „Rozumiem, że celem Międzynarodowej Korporacji Finansowej jest wykazanie w konkretnej formie, że inwestycje rozsądnie kierowane do słabiej rozwiniętych regionów mogą być wysoce zyskowne i tym mogą za-