

LXIII Zjazd PTJ, Warszawa

16-17.09.2003

Pomor, Humor

Morfeusz SIAT

Poliqarp

Holmes

Kryteria wyboru

Robert Wołosz

Marcin Woliński

Adam Przepiórkowski

Michał Rudolf

Niebieska gramatyka

Saloni, Świdziński

Marcin Woliński

Czy to jest ważne?

Czy to jest ważne?

Robert Wołosz

Robert Wołosz

Robert Wołosz

Adam Przepiórkowski

Adam Przepiórkowski

Marcin Woliński

Robert Wołosz

Robert Wołosz

Robert Wołosz

Robert Wołosz

Robert Wołosz

Robert Wołosz

Marcin Woliński

Michał Rudolf

Michał Rudolf

Status pojęcia słowa

Słowo a wyraz

Słowo w potocznym rozumieniu

Marcin Woliński

Marcin Woliński

Adam Przepiórkowski

Michał Rudolf

Pisownia łączna i rozłączna

Szkoła Tokarskiego

Marcin Woliński

Adam Przepiórkowski

Korpus IPI PAN

Inne pojęcia

LXIII Zjazd PTJ, Warszawa

16-17.09.2003

JANUSZ S. BIEŃ

Aparat pojęciowy
wybranych systemów
przetwarzania tekstów polskich

17.09.2005

ROBERT WOŁOSZ

**Efektywna metoda analizy i syntezy
morfologicznej w języku polskim.**

Niepublikowana praca doktorska.

Wydział Polonistyki,

Uniwersytet Warszawski,

Warszawa 2000.

MARCIN WOLIŃSKI.

**Komputerowa weryfikacja
gramatyki Świdzińskiego.**

Niepublikowana praca doktorska.

Instytut Podstaw Informatyki PAN,

Warszawa 2004.

<http://nlp.ipipan.waw.pl/~wolinski/morfeusz/>

ADAM PRZEPIÓRKOWSKI.

Korpus IPi PAN. Wersja wstępna.

Instytut Podstaw Informatyki PAN

Warszawa 2004.

<http://korpus.pl>

MICHAŁ RUDOLF.

**Metody automatycznej analizy
korpusu tekstów polskich:
pozyskiwanie, wzbogacanie i
przetwarzanie informacji
lingwistycznych.**

Praca doktorska.

Wydział Polonistyki,
Uniwersytet Warszawski,
Warszawa 2003.

MICHAŁ RUDOLF.

**Metody automatycznej analizy
korpusu tekstów polskich:
pozyskiwanie, wzbogacanie i
przetwarzanie informacji
lingwistycznych.**

Wydział Polonistyki,
Uniwersytet Warszawski,
Warszawa 2004[2005].

Aparat pojęciowy
wybranych systemów
przetwarzania tekstów polskich

Rozumienie terminów *słowo*, *forma hasłowa*, *forma wyrazowa* i *leksem* przejmuję z pracy: Saloni Świdziński, 1998.

Punktem wyjścia dla dalszych
rozważań będzie system pojęciowy
*Składni współczesnego języka
polskiego.*

Wiele rozwiązań opisanych
w niniejszym rozdziale zostało
zaczerpniętych z prac Zygmunta
Saloniego i jego współpracowników

W niniejszej pracy przyjmuję
w większości opis fleksji i składni
polskiej przedstawiony w *Składni
współczesnego języka polskiego*



Ciąg liter pomiędzy sąsiednimi
spacjami będziemy nazywać *słowem*.

Definicja słowa, choć przejrzysta, nie odpowiada w pełni intuicji językowej

<http://korpus.pwn.pl/>

Wydawnictwo Naukowe PWN
przygotowało i udostępniło sieciową
wersję Korpusu Języka Polskiego
PWN wielkości 34,5 miliona słów.

[http://pelcra.ia.uni.lodz.pl/
corpora_pl.php](http://pelcra.ia.uni.lodz.pl/corpora_pl.php)

Część zbioru (obecnie - 30 milionów słów) zostało już opracowane jako zbilansowany korpus

Przedmiotem analizy są słowa graficzne, rozumiane jako ciągi liter między dwoma spacjami lub znakami o wartości spacji.

Do alfabetu wejdą pewne znaki na
prawach liter

Nie zaliczymy do liter znaków przestankowych ani znaków graficznych. Należy jednak uczynić wyjątek dla kropki w ustabilizowanych skrótach.

słowa rozumiane jako maksymalne ciągi znaków nie będących separatorami słów,

separatorami słów są odstępy oraz znaki interpunkcyjne z wyłączeniem dywizu, kropki będącej częścią skrótów oraz apostrofu w formach takich jak *Chomsky'ego* i *(de) l'Hospitala*.

Tymczasem wydaje się wygodne uznanie napisu *Lagrange'a* za jedno słowo. Dotyczy to również napisów takich jak *ping-pong* i *PRL-u*.
Horthy'ego

Do alfabetu wejdą pewne znaki na prawach liter:

a) apostrof - w języku polskim najczęściej w środku słowa, por. dell'arte (SJPDor.), Horthy'ego

Do alfabetu wejdą pewne znaki na
prawach liter: [...]

d) łącznik, czyli dywiz [...]

d) łącznik, czyli dywiz (czasami pojawia się on w sposób ustabilizowany w słowach, których części nie funkcjonują samodzielnie, por. tse-tse, cza-cza, tam-tamista (SJPDor.) - inaczej niż mający wyraźne cechy samodzielnego słowa człon polsko- w złożeniach typu polsko-radziecki [...]).

Do alfabetu wejdą pewne znaki na
prawach liter: [...]

b) cyfry - pisane łącznie
z tradycyjnymi literami alfabetu, por.
126p, F-16;

Do alfabetu wejdą pewne znaki na
prawach liter: [...]

c) ukośnik (kreska ukośna: /) - pisany
z tradycyjnymi literami alfabetu, por.
m/s (SPP);

Nie zaliczymy do liter znaków przestankowych ani znaków graficznych. Należy jednak uczynić wyjątek dla kropki w ustabilizowanych skrótach.

Kolejnym problematycznym znakiem jest kropka, która występuje w tekście w dwóch funkcjach: jako znak interpunkcyjny oraz jako obowiązkowa część skrótu. W tym drugim wypadku kropkę traktuję jako część słowa.

Napisem nazywać będę dowolną sekwencją znaków (liter, cyfr), być może zawierającą dywiz, która stanowi samodzielny fragment tekstu, to jest zarówno przed nią, jak i po niej znajduje się spacja, znak interpunkcyjny lub granica wypowiedzenia. Oprócz tego napisem jest każdy znak interpunkcyjny.

Słowo to napis nie będący znakiem interpunkcyjnym, interpretowany bez uwzględniania kontekstu.

Słowo:
unilateralne
czy
bilateralne?

Słowo: *unilateralne*

Wyraz: *bilateralny*

Słowo:

opłata za telegram

opłata za ogłoszenie drobne

[...]

podwójne pstryknięcie myszą

Znaki interpunkcyjne są traktowane jako pełnoprawne składniki wypowiedzenia, w związku z czym ich brak powoduje, że wypowiedzenie staje się nieakceptowane przez gramatykę.

Dla konsekwencji również segmenty
złożone ze znaków interpunkcyjnych
[...]

Znaki interpunkcyjne będące separatorami słów traktowane są jako osobne segmenty. [...]

Oprócz tego napisem jest każdy znak
interpunkcyjny.

Czyś pisał?

Czy pisałeś?

Czy+ś pisał?

Czy pisał+eś?

Miałem miał.

Ile interpretacji?

interpretowalne ciągi znaków
nazywać będziemy *segmentami* [...]

Segmenty: 360 446 336

Słowa: 291 187 457

Współczynnik: 1,24%

- fleksemy
- kubliki
- burkinostki
- . . .

Dziękuję za uwagę!

Proszę o pytania.