



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Praca Magisterska

**Automatyczna kontekstowa korekta tekstów
na podstawie Grafu Przyzwyczajień
Lingwistycznych zbudowanego przez robota
internetowego dla języka polskiego**

Marcin A. Gadamer

**Promotor:
dr Adrian Horzyk**

Agenda

- 1. Wstęp**
- 2. Cel pracy**
- 3. Rozwiązanie**
 - **Pająk internetowy**
 - **Konstrukcja grafu LHG**
 - **Algorytm do korekcji tekstu**
 - **Sprawdzanie poprawności wprowadzanych słów**
 - **Podpowiadanie dokończania słów**
 - **Automatyczna kontekstowa korekta tekstów**
- 4. Porównanie otrzymanych wyników z**
 - **Microsoft Word 2007**
 - **OpenOffice.org Writer 3**
 - **Wyszukiwarką Google.com**
- 5. Wnioski**
- 6. Zakończenie**



AGH

Wstęp

- Można dziś łatwo zauważyć ewolucję, jakiej uległ sposób komunikacji międzyludzkiej oraz związana z nim przemiana języka.
- W dzisiejszych czasach powstaje bardzo duża liczba elektronicznych dokumentów, w których nieodłącznie występują błędy językowe.
- Istnieje więc potrzeba rozumienia tekstów oraz stworzenia bardziej inteligentnej i kontekstowej korekty tekstu, który został wprowadzony z różnego rodzaju błędami.

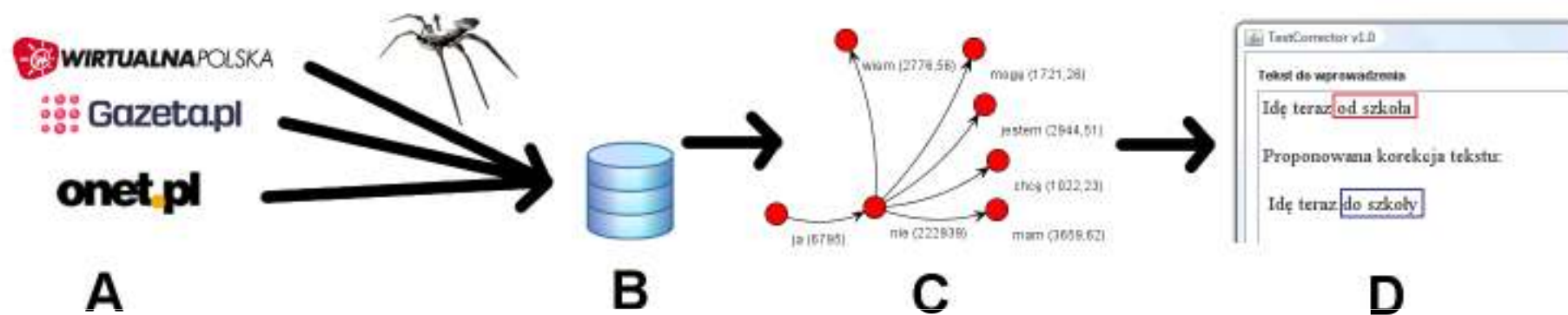
Celem niniejszej pracy było skonstruowanie i przetestowanie algorytmu automatycznej kontekstowej korekty tekstu z wykorzystaniem grafu LHG, który umożliwia badanie kontekstu w wymiarze częstotliwościowo-fleksyjnym.

Cele szczegółowe pracy

Zakres pracy obejmował w szczególności:

- zbudowanie specjalistycznego robota (pająka) internetowego,
- zbudowanie maksymalnie rozbudowanego i ogólnego Grafu Przyzwyczajzeń Lingwistycznych (LHG),
- skonstruowanie algorytmu, służącego do automatycznego wykrywania błędów językowych i ich poprawy na podstawie kontekstu,
- przetestowanie stworzonej aplikacji i porównanie wyników otrzymanej korekty różnych tekstów z innymi programami.

Schemat działania algorytmu:





AGH

Rozwiązanie - opis

- Specjalistyczny pajak internetowy zbiera informacje o zdaniach ze stron internetowych i zapisuje je w bazie danych.
- Dzięki analizie zebranych korpusów tekstów w bazie danych tworzony jest Graf Przyzwyczajień Lingwistycznych (LHG) dla języka polskiego.
- Na podstawie grafu LHG innowacyjny algorytm dokonuje automatycznej kontekstowej korekty wprowadzonych tekstów bazując na kontekście słowno-fleksyjnym i częstotliwości jego występowania oraz podobieństwie słów występujących w tym kontekście do słów błędnych.



AGH

Pająk internetowy - opis

Specjalistyczny pająk internetowy zebrał informacje o zdaniach ze stron internetowych i zapisał je w bazie danych.

Adres sprawdzanej strony WWW

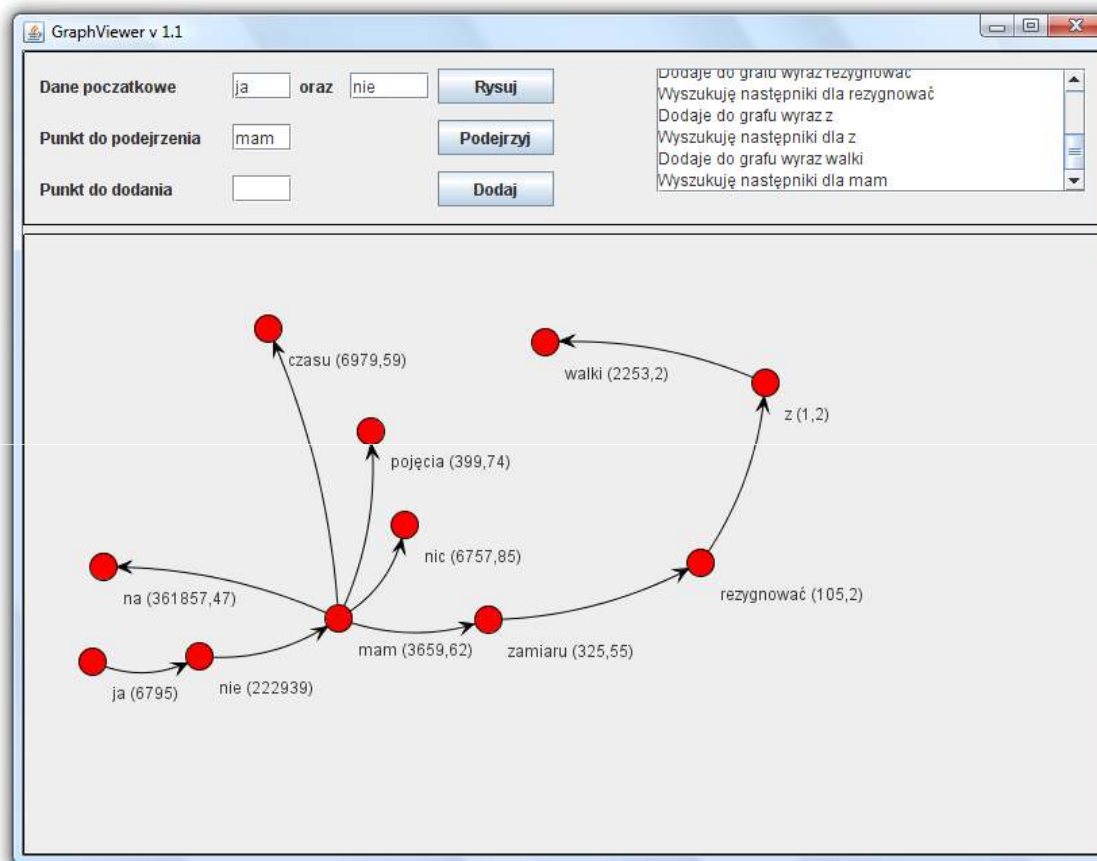
```
2 heartBip! Tue May 19 08:38:28 CEST 2009 %stat% url: http://wyborcza.pl/0,0.html  
pkt - start: 3130 finish: 153056 href size: 676 wordBaseModel: 2176
```

Ilość znalezionych odnośników
do kolejnych stron

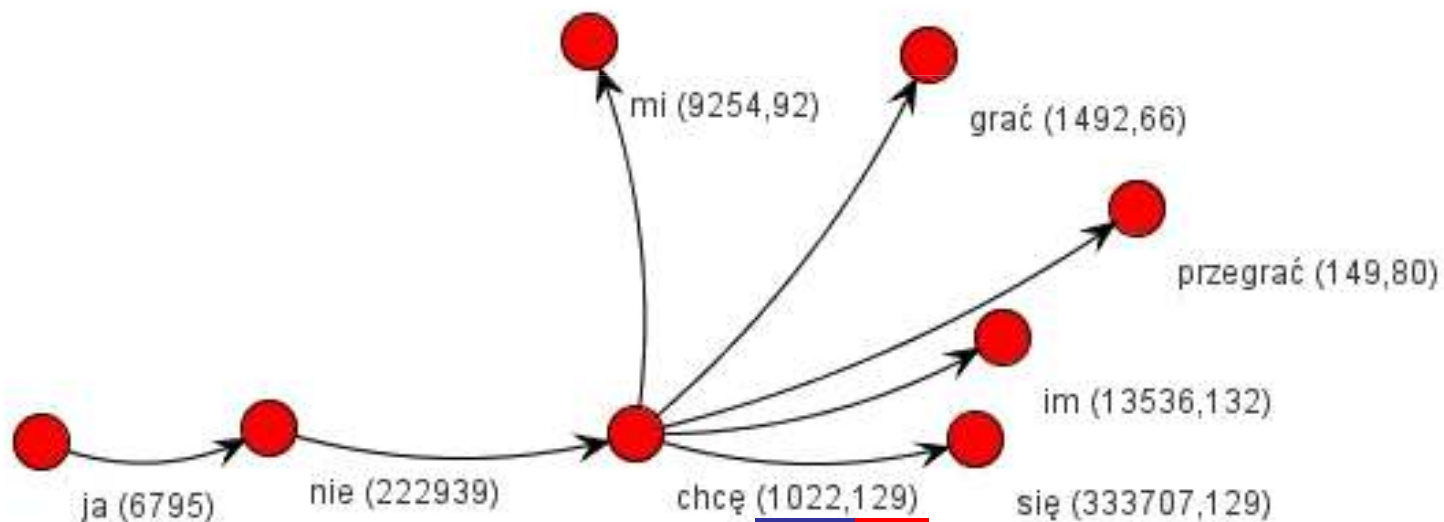
Ilość dodanych słów-trójek
do bazy danych

Konstrukcja grafu LHG

Dzięki analizie zebranych korpusów tekstów w bazie danych stworzony został Graf Przyzwyczajzeń Lingwistycznych (LHG) dla języka polskiego.



Opis skonstruowanego grafu LHG - symbolika i częstotliwość słów



Częstość słowa „chcę”
w Słowniku frekwencyjnym

Częstość słowa „chcę”
w określonym kontekście

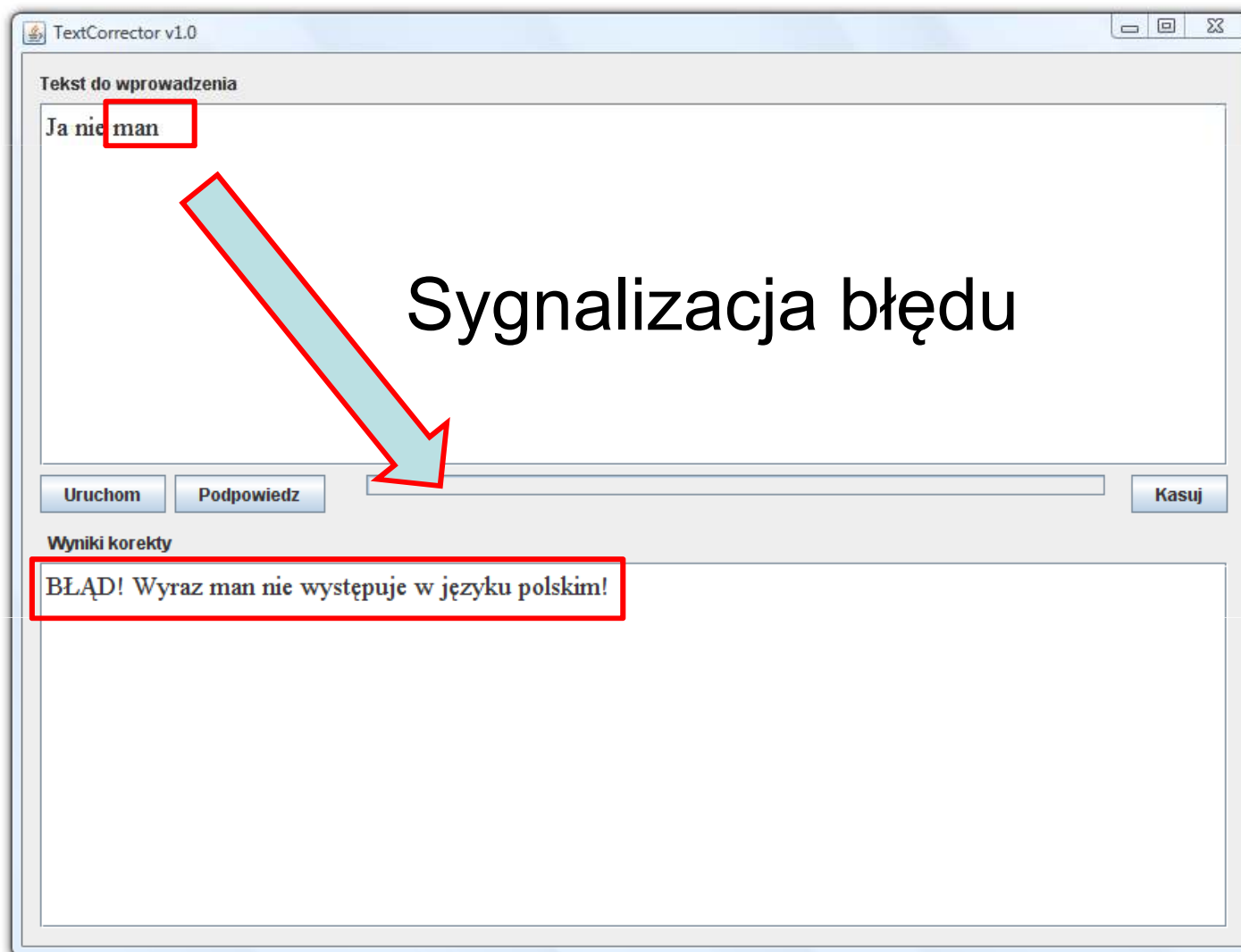


AGH

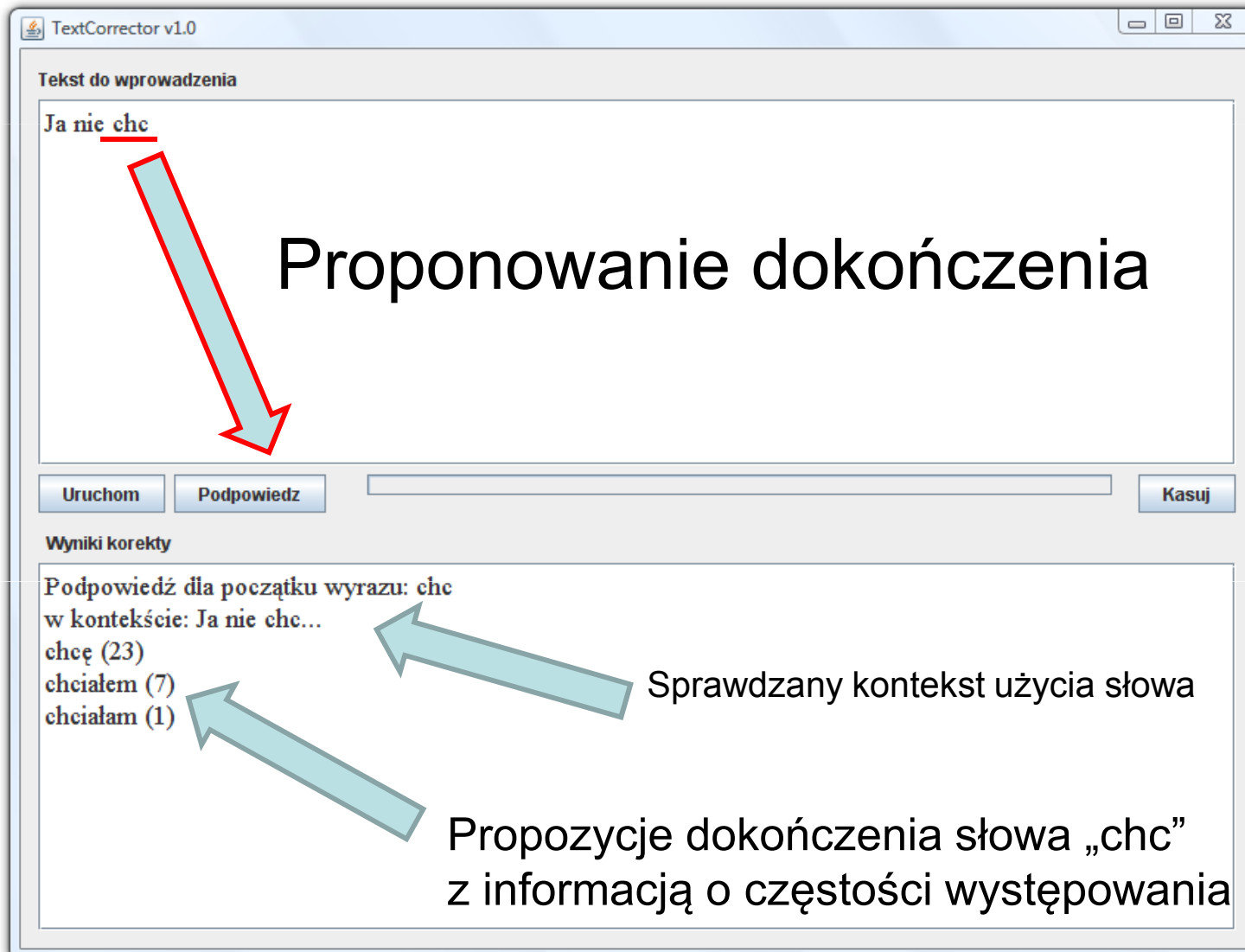
Idea rozwiązania

Na podstawie grafu LHG innowacyjny algorytm dokonuje automatycznej kontekstowej korekty wprowadzonych tekstów bazując na kontekście słowno-fleksyjnym i częstotliwości jego występowania oraz podobieństwie słów występujących w tym kontekście do słów błędnych.

Sprawdzanie poprawności wprowadzanych słów



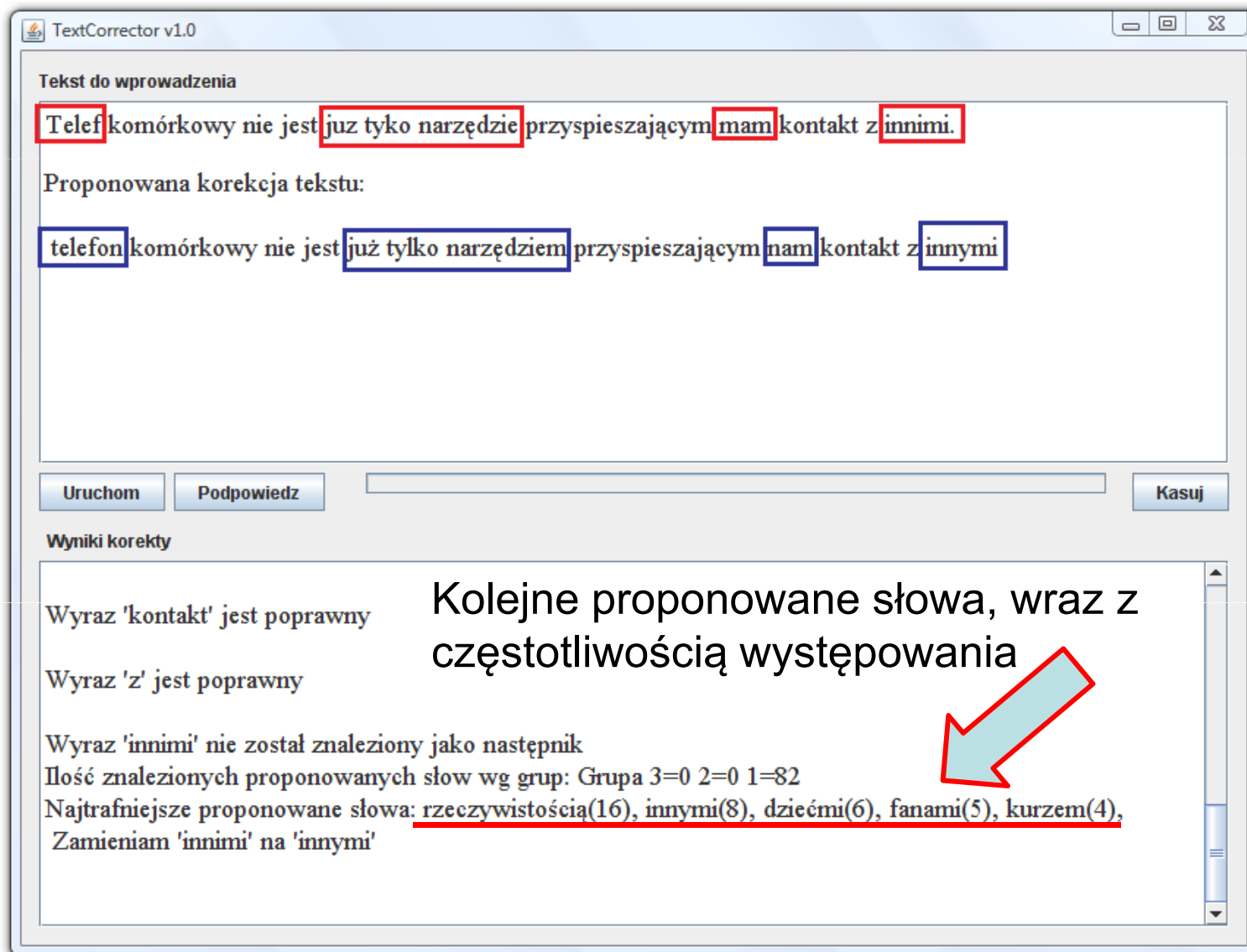
Podpowiadanie dokończenia słów



The screenshot shows a software window titled "TextCorrector v1.0". The main text area contains "Ja nie chc" with "chc" underlined. A red arrow points from the underlined text to the title "Proponowanie dokończenia". Below the text area are buttons for "Uruchom", "Podpowiedz", and "Kasuj". The "Wyniki korekty" section displays the following information:

- Podpowiedź dla początku wyrazu: chc
- w kontekście: Ja nie chc...
- chcę (23)
- chciałem (7)
- chciałam (1)

Two blue arrows point from the list of suggestions to the text "Sprawdzany kontekst użycia słowa" and "Propozycje dokończenia słowa „chc” z informacją o częstości występowania".



TextCorrector v1.0

Tekst do wprowadzenia

Telef komórkowy nie jest już tylko narzędzie przyspieszającym mam kontakt z innymi.

Proponowana korekcja tekstu:

telefon komórkowy nie jest już tylko narzędziem przyspieszającym nam kontakt z innymi

Uruchom Podpowiedz Kasuj

Wyniki korekty

Wyraz 'kontakt' jest poprawny

Wyraz 'z' jest poprawny


Wyraz 'innimi' nie został znaleziony jako następnik

Ilość znalezionych proponowanych słów wg grup: Grupa 3=0 2=0 1=82

Najtrafniejsze proponowane słowa: rzeczywistością(16), innymi(8), dziećmi(6), fanami(5), kurzem(4),

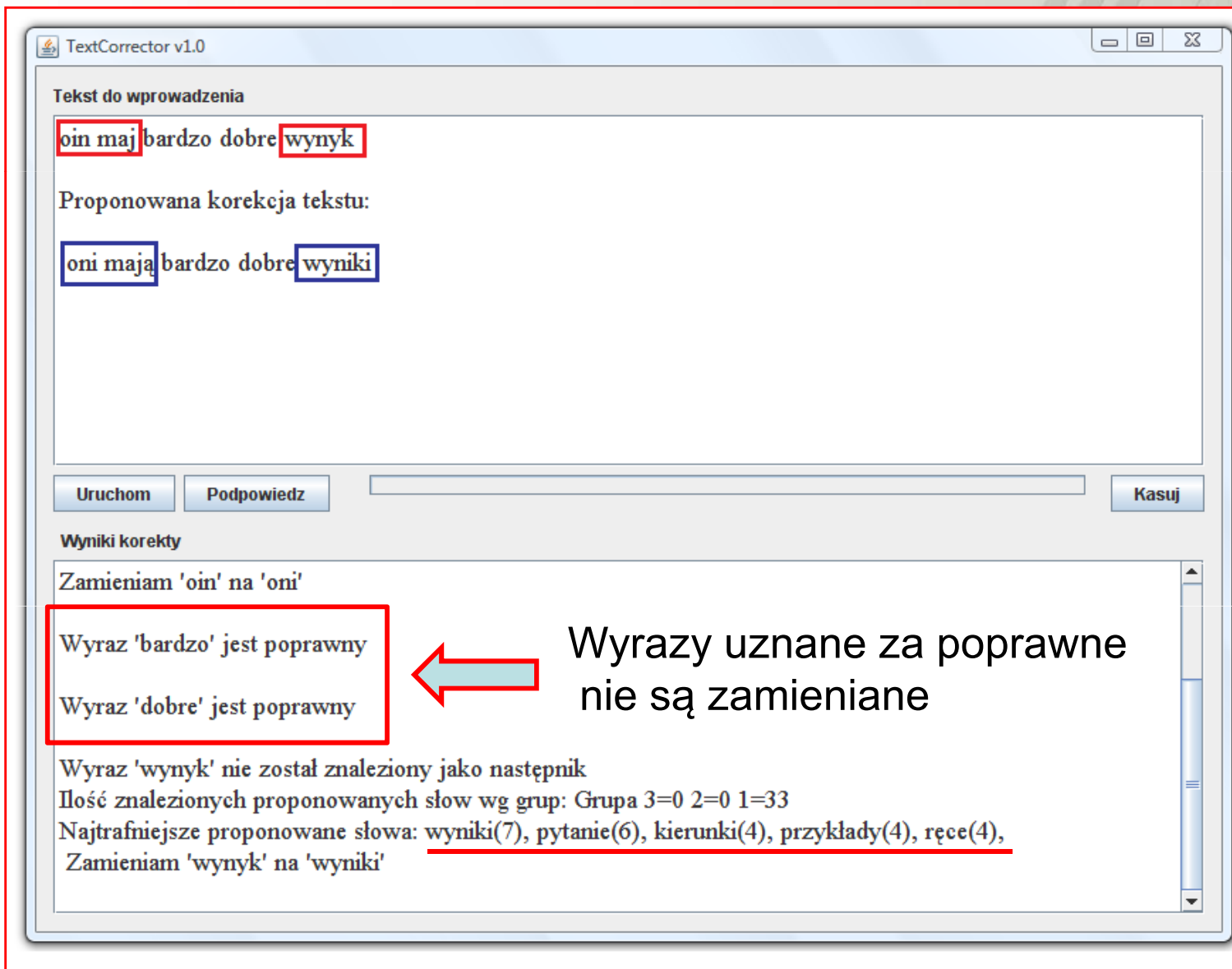
Zamieniam 'innimi' na 'innymi'

Kolejne proponowane słowa, wraz z częstotliwością występowania



Rozwiązanie – główna aplikacja

Automatyczna kontekstowa korekta tekstów



The screenshot shows the TextCorrector v1.0 application window. The main text area contains the input text "oin maj bardzo dobre wynik" with "oin maj" and "wynik" highlighted in red boxes. Below it, the suggested correction is "oni mają bardzo dobre wyniki" with "oni mają" and "wyniki" highlighted in blue boxes. The "Wyniki korekty" section shows the following results:

- Zamieniam 'oin' na 'oni'
- Wyraz 'bardzo' jest poprawny
- Wyraz 'dobre' jest poprawny
- Wyraz 'wynyk' nie został znaleziony jako następnik
- Ilość znalezionych proponowanych słów wg grup: Grupa 3=0 2=0 1=33
- Najtrafniejsze proponowane słowa: wyniki(7), pytanie(6), kierunki(4), przykłady(4), ręce(4)
- Zamieniam 'wynyk' na 'wyniki'

A red arrow points from the text "Wyrazy uznane za poprawne nie są zamieniane" to the two lines "Wyraz 'bardzo' jest poprawny" and "Wyraz 'dobre' jest poprawny".


Porównanie z innymi aplikacjami

Idę teraz od szkoła

Brak podpowiedzi

Microsoft Word 2007

Idę teraz od szkoła

 Przyimek wymaga dopełniacza

Open Office Writer 3.0



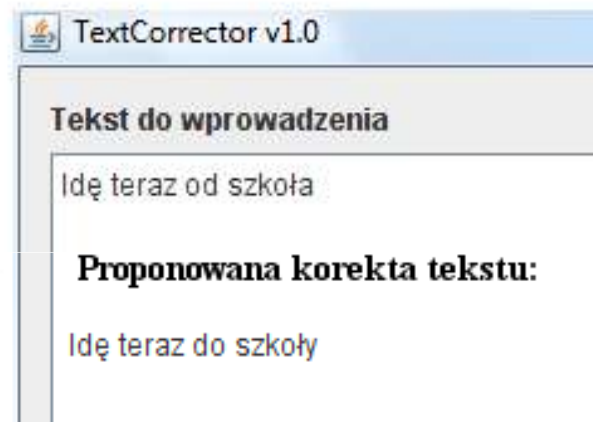
Idę teraz od szkoła

Szukaj w internecie

Sieć

Brak podpowiedzi

wyszukiwarka Google



TextCorrector v1.0

Tekst do wprowadzenia

Idę teraz od szkoła

Proponowana korekta tekstu:

Idę teraz do szkoły

stworzona aplikacja

Proponowane korekty wyrazów w zdaniu „Oin maj bardzo dobre wynyk”

- Microsoft Word 2007
 - *oin* – sugerowana poprawa: Ion, ino, oni, on
 - *wynyk* – sugerowana poprawa: wnyk, wymyk, wytyk, wynik
- OpenOffice.org Writer 3.0
 - *oin* – sugerowana poprawa: In, Zin, Ocin, Osin
 - *wynyk* – sugerowana poprawa: wnyk, wytyk, wymyk, wynik
- wyszukiwarka Google
 - *oin* – sugerowana poprawa: oni
 - *maj* – sugerowana poprawa: mają
 - *wynyk* – sugerowana poprawa: wyniki
- Stworzona aplikacja
 - *oin* – sugerowana poprawa: oni
 - *maj* – sugerowana poprawa: mają
 - *wynyk* – sugerowana poprawa: wyniki

Proponowane korekty wyrazów w zdaniu „Oni mają bardo madre dzieci”

- Microsoft Word 2007
 - *madre* – sugerowana poprawa: Madre
 - *dzieic* – sugerowana poprawa: dzieci, Dzielic, dziewic, Dzibic
- OpenOffice.org Writer 3.0
 - *bardo* – sugestia prawdopodobnej literówki
 - *madre* – sugerowana poprawa: mądre, mader, Mare, made
 - *dzieic* – sugerowana poprawa: dzieci, dziec, dziewic
- wyszukiwarka Google
 - *bardo* – sugerowana poprawa: bardzo
 - *dzieic* – sugerowana poprawa: dzieci
- Stworzona aplikacja
 - *bardo* – sugerowana poprawa: bardzo
 - *madre* – sugerowana poprawa: mądre
 - *dzieic* – sugerowana poprawa: dzieci

Jak zostało wykazane, skonstruowany mechanizm kontekstowej korekty okazał się być znacznie bardziej skuteczny od konkurencji w rozpoznawaniu miejsc, w których wystąpiły błędy oraz w ich korekcie.



Zalety innowacyjnego algorytmu

Główne zalety zaproponowanego algorytmu:

- Moduł do tworzenia zdań.
- Innowacyjny sposób wykrywania potencjalnych błędów w tekście.
- Inteligentne dopełnianie wyrazów.
- Automatyczna kontekstowa korekta tekstów.

Graf LHG w połączeniu z zaproponowanym algorytmem korekty tekstów dał możliwość skonstruowania bardziej efektywnych aplikacji do automatycznej korekty tekstów dla języka polskiego, ze względu na wykorzystanie odpowiednio skompresowanego kontekstu słowno-fleksyjnego z uwzględnieniem częstotliwości jego występowania w języku polskim.

Dziękuję za uwagę