

Spis treści

Przedmowa	7
<i>Ryszard Tadeusiewicz</i>	
Wstęp	11
<i>Wiesław Lubaszewski</i>	
I. SŁOWNIKI KOMPUTEROWE	
1. Wyraz	15
<i>Wiesław Lubaszewski</i>	
1.1. Wyraz w słowniku fleksyjnym	17
1.2. Wyrazy wielosegmentowe	23
1.3. Potencjalne wyrazy wielosegmentowe	25
1.4. Znaczenie wyrazu	26
2. Słownik fleksyjny	37
<i>Piotr Pisarek</i>	
2.1. System klasyfikacji fleksyjnej wyrazów języka polskiego	37
2.2. System generowania słownika fleksyjnego	47
2.2.1. Gramatyki lokalne	48
2.2.2. Rozpoznawanie gramatyk lokalnych – wzorców odmiany	59
2.2.3. Test skuteczności wygenerowanego słownika	64
3. Słownik wyrazów wielosegmentowych	69
<i>Michał Rokitiński</i>	
3.1. Leksykon	69
3.2. Algorytm rozpoznawania wyrazów wielosegmentowych	72
4. Reprezentacja wielosegmentowych wyrazów potencjalnych	79
<i>Michał Korzycki</i>	
4.1. Model relacyjny i jego notacja	79
4.2. Relacyjny model przymiotników potencjalnych	82
4.3. Relacyjny model liczebników	84
4.3.1. Model relacyjny wielowyrzowych liczebników głównych	85
4.3.2. Model relacyjny wielowyrzowych liczebników porządkowych	88
4.3.3. Postać graficzna liczebnika w tekście	90

4.3.4. Wartość semantyczna liczebnika.....	92
4.3.5. Postać znacznika dla liczebników.....	93
4.4. Podsumowanie	93
5. Wielowarstwowy model słownika i rozpoznawanie wyrazu w tekście	95
<i>Michał Korzycki, Wiesław Lubaszewski</i>	
5.1. Rozpoznawanie wyrazu w tekście	95
5.2. Warstwowy system rozpoznawania wyrazu	97
5.3. Rozstrzyganie wieloznaczności za pomocą relacyjnego opisu znaczenia	102
5.4. Rozpoznawanie form dyskretnych.....	104
5.5. Interpretacja nierozpoznanych form tekstowych.....	105
6. Słownik fleksyjny jako biblioteka języka C	107
<i>Marek Gajęcki</i>	
6.1. CLP – warstwa fleksyjna	107
6.1.1. Baza wyrazów pierwotnych (słownik fleksyjny).....	107
6.2. Interfejs bazy wyrazów pierwotnych.....	113
6.3. CLP – warstwa morfologiczna.....	116
6.3.1. Baza relacji morfologicznych w bibliotece CLP	117
6.3.2. Interfejs bazy relacji morfologicznych	118
6.4. Przykłady zastosowań biblioteki CLP z interfejsem ILP	118
6.4.1. Rozpoznawanie form tekstowych	118
6.4.2. Eksport zawartości słownika fleksyjnego	119
6.4.3. Program typu grep.....	120
6.4.4. Statystyka wyrazów w zbiorze tekstów – prawo Zipfa	121
6.4.5. Wektorowa postać tekstu, teksty podobne	124
6.4.6. Kolokacje	126

II. EKSTRAKCJA INFORMACJI Z TEKSTU

7. Informacja w tekście.....	137
<i>Wiesław Lubaszewski</i>	
7.1. Operacje na wyrazach.....	137
7.2. Informacja niesiona przez tekst	139
7.3. Informacja tekstowa a operacje na wyrazach	139
7.4. Informacja skojarzona i granice dopasowania.....	142
7.5. Schemat informacyjny przygotowany ręcznie.....	143
7.6. Schemat informacyjny wyekstrahowany z tekstu	146
7.7. Dewiacja informacyjna i informacja krytyczna.....	151
8. Warstwowa ekstrakcja informacji z tekstu na przykładzie systemu automatycznej oceny CV	155
<i>Jadwiga Babińska, Agnieszka Doniec</i>	
8.1. Reprezentacja wiedzy	155
8.2. Ekstrakcja informacji z tekstu	157
8.2.1. Identyfikacja wyrazu.....	157
8.2.2. Identyfikacja wzorca	157

8.2.3.	Generowanie oczekiwań	158
8.2.4.	Weryfikacja oczekiwań (<i>expectation-matching</i>).....	158
8.3.	Reinterpretacja wyekstrahowanych informacji.....	159
8.3.1.	Ocena ogólna cv.....	159
8.3.2.	Reinterpretacja oceny przez profil pracodawcy.....	161
8.4.	Podsumowanie	162
8.5.	Dodatek 1	162
8.6.	Dodatek 2	163
9.	Tekst jako wzorzec informacyjny – automatyczna ocena podobieństwa tematycznego tekstów za pomocą <i>Latent Semantic Analysis</i>	165
	<i>Agnieszka Figiel</i>	
9.1.	Przygotowywanie danych	166
9.2.	Optymalizacja słownika frekwencyjnego	166
9.3.	Dobór schematu wagowego	167
9.4.	Redukcja wymiarów w LSA.....	168
9.5.	Wybór ilości wymiarów LSA	169
9.6.	Dołączanie nowego tekstu	169
9.7.	Ocena podobieństwa tekstów.....	170
9.8.	Ocena skuteczności algorytmu	170
9.8.1.	Precyzja i przywołanie	171
9.8.2.	Wynik wzorcowy	172
9.8.3.	Umotywowane odstępstwa od wzorca.....	174
9.8.4.	Porównanie uśrednionej interpolowanej precyzji.....	174
9.8.5.	Porównanie wpływu podzbioru badanych wyrazów	175
9.8.6.	Porównanie wpływu schematu wagowego	176
9.9.	Wnioski	177
10.	System automatycznej ekstrakcji tekstów polskojęzycznych z Internetu	179
	<i>Krzysztof Dorosz</i>	
10.1.	Wprowadzenie.....	179
10.1.1.	<i>Web crawling</i>	179
10.1.2.	Crawling wyspecjalizowany	180
10.2.	Crawling wyspecjalizowany w ekstrakcji tekstów polskojęzycznych	181
10.2.1.	Architektura.....	181
10.2.2.	<i>HTML Stripping</i>	183
10.2.3.	Rozpoznanie języka tekstu.....	184
10.2.4.	Implemetacja	185
10.3.	Zastosowania.....	185
11.	Ekstrakcja leksykalna	187
	<i>Paweł Pietras</i>	
11.1.	Wprowadzenie.....	187
11.2.	Uogólniony opis systemu ekstrakcji nazw	189
11.3.	System automatycznego rozpoznawania typu fleksyjnego na podstawie dowolnej formy tekstowej	192

11.4. System bezkontekstowej kategoryzacji semantycznej nazw jednosegmentowych	201
11.4.1. Konstrukcja systemu reguł klasyfikacji semantycznej	208
11.4.2. System automatycznej generacji reguł klasyfikacji semantycznej ...	215
11.5. Implementacja algorytmów	224
11.6. Testy skuteczności	224
11.6.1. Przygotowanie testów	225
11.6.2. Opracowanie wyników testów	229
12. Rozstrzygnięcie wieloznaczności, maszynowa reprezentacja znaczenia wyrazu i ekstrakcja znaczeń	241
<i>Aleksander Pohl</i>	
12.1. Wieloznaczność	241
12.2. Słownik Semantyczny Języka Polskiego	243
12.3. Ontologia CYC	249
12.4. Podsumowanie	255
Bibliografia	257