

# Spis treści

Spis treści . . . . .	5
Wstęp . . . . .	9
1. Komputerowe przetwarzanie tekstów języka polskiego . . . . .	11
1.1. Przegląd komputerowych narzędzi przetwarzania języka polskiego . . . . .	12
1.1.1. Podział na zdania . . . . .	12
1.1.2. Tokenizacja . . . . .	16
1.1.3. Analiza morfologiczna . . . . .	29
1.1.4. Ujednoznacznianie morfosyntaktyczne . . . . .	45
1.1.5. Parsing płytki . . . . .	51
1.1.6. Rozpoznawanie jednostek nazewniczych . . . . .	60
1.1.7. Parsing głęboki . . . . .	71
1.1.8. Normalizacja tekstu . . . . .	84
1.2. Przegląd polskich systemów integrujących narzędzia przetwarzania języka . . . . .	90
1.2.1. Language Tool . . . . .	90
1.2.2. Multiservice . . . . .	93
1.2.3. CLARIN . . . . .	97
2. PSI-Toolkit – zintegrowany, kompatybilny i rozszerzalny zestaw narzędzi przetwarzania języka naturalnego . . . . .	99
2.1. Przeznaczenie systemu PSI-Toolkit . . . . .	99
2.2. Struktura danych (krata) . . . . .	101
2.2.1. Anotacja krawędzi . . . . .	103
2.2.2. Tagi krawędzi . . . . .	103
2.2.3. Rozkłady na krawędzie składowe . . . . .	103
2.2.4. Wagi krawędzi . . . . .	103
2.2.5. Motywacja zdefiniowania kraty jako struktury danych . . . . .	104
2.2.6. Tekstowa postać <i>PSI-kraty</i> . . . . .	105
Podsumowanie podrozdziału 2.2 . . . . .	105
2.3. Potoki przetwarzania . . . . .	106
Podsumowanie podrozdziału 2.3. . . . .	106
2.4. Procesory . . . . .	106
2.4.1. Readery . . . . .	107
2.4.2. Anotatory zestawu PSI-Toolkit . . . . .	111
2.4.3. Writery . . . . .	119
Podsumowanie podrozdziału 2.4 . . . . .	121

2.5. Odseparowanie danych językowych od procesu przetwarzania . . . . .	121
2.6. Rozszerzalność zestawu <i>PSI-Toolkit</i> . . . . .	122
2.7. Przenośność . . . . .	122
3. Normalizacja diachroniczna tekstów języka polskiego . . . . .	125
3.1. Motywacja badań . . . . .	125
Podsumowanie podrozdziału 3.1. . . . .	129
3.2. Przegląd komputerowych systemów normalizacji tekstu . . . . .	129
3.2.1. Metody normalizacji tekstu . . . . .	129
3.2.2. Prace związane z normalizacją diachroniczną . . . . .	132
Podsumowanie podrozdziału 3.2 . . . . .	135
3.3. <i>iaiko</i> – program do normalizacji diachronicznej z wykorzystaniem trans-	
duktorów skończenie stanowych . . . . .	135
3.3.1. Początkowe źródło reguł normalizacji . . . . .	135
3.3.2. Transduktory skończenie-stanowe . . . . .	136
3.3.3. Przykłady reguł normalizacji diachronicznej w systemie <i>iaiko</i> . . . . .	137
3.3.4. Lista wyjątków . . . . .	144
3.3.5. Wybrane aspekty implementacji . . . . .	145
3.3.6. Ewaluacja rozwiązania . . . . .	147
Podsumowanie podrozdziału 3.3 . . . . .	152
3.4. <i>Diachronic</i> normalizer – narzędzie do normalizacji diachronicznej jako ele-	
ment zestawu <i>PSI-Toolkit</i> . . . . .	152
3.4.1. Normalizacja diachroniczna w procesie przetwarzania tekstu . . . . .	152
3.4.2. Przykłady użycia procesora <i>iaiko</i> . . . . .	153
3.4.3. Opóźnione ujednoznacznianie wyników normalizacji diachronicznej	155
3.4.4. Tworzenie spersonalizowanego normalizatora . . . . .	158
Podsumowanie podrozdziału 3.4. . . . .	159
4. Reguły normalizacji diachronicznej . . . . .	160
4.1. „Odkrywka” – korpusowe źródło datowania reguł . . . . .	160
4.2. Metodyka datowania reguł . . . . .	161
4.3. Reguły inspirowane opracowaniem Felińskiego (1816) . . . . .	162
Reguła „i -> j” po „y”, przed spółgłoską . . . . .	163
Reguła „i -> j” na początku wyrazu przed samogłoską . . . . .	164
Reguła „i -> j” pomiędzy samogłoskami . . . . .	165
Reguła „i->j” po samogłosce przed spółgłoską . . . . .	166
Reguła „y -> j” przed samogłoską . . . . .	167
Reguła usunięcia litery „y” . . . . .	168
Reguła „y -> j” po samogłosce . . . . .	169
Reguła „dź-> ć” w bezokolicznikach . . . . .	170
Reguła „dz-> c” w bezokolicznikach . . . . .	171
Reguła „z -> s” w przedrostkach . . . . .	173
Reguła „á -> a” . . . . .	174
Reguła „iej -> ij” w formach rozkazujących . . . . .	176
Reguła „s -> z” w przyimkach. . . . .	177

4.4. Reguły inspirowane Rozprawami Deputacji (Rozprawy, 1830)	178
Reguła “szy -> łszy” w formach imiesłowów przysłówkowych uprzednich	178
Reguła „źródło”	179
Reguła: polonizacja nazw własnych	181
Reguła “gie -> ge” w wyrazach obcego pochodzenia	182
Reguła “ó -> o” w wybranych wyrazach	183
Rule: “x-> ks”	185
Reguła modernizująca przymiotniki rodzaju męskiego w narzędniku	186
4.5. Reguły inspirowane Uchwałami Akademii Umiejętności (1891)	187
Reguła “zk -> sk” w przymiotnikach	187
Reguła “z -> s” w końcówkach rzeczowników “ęstwo”	188
Reguła “z -> s” w przedrostku	189
Reguła „é -> e”	190
4.6. Reguły inspirowane Zasadami Akademii Umiejętności (Zasady, 1917)	192
Reguła z -> ś w przedrostkach czasowników	192
4.7. Reguły inspirowane Uchwałami Komitetu Ortograficznego (Jodłowski, Taszycki, 1936)	193
Reguła yjum -> ium w zakończeniach rzeczowników	194
Reguła yjum -> jum w zakończeniach rzeczowników	195
Reguła yja -> ia w zakończeniach rzeczowników	196
Reguła yja -> ja w zakończeniach rzeczowników	197
Reguła yj -> ii w zakończeniach rzeczowników	198
Reguła yj -> ji w zakończeniach rzeczowników	199
Reguła y -> i po spółgłosce, a przed samogłoską	200
4.8. Reguły podziału wyrazów	201
Reguła “nie” + czasownik	201
Reguły podziału specyficznych wyrazów	202
Reguły podziału spójników złożonych typu “przyczem”	203
Reguła odseparowania partykuły “by” od wyrazów nieodmiennych	204
Reguła rozdzielenia partykuły “nie” od predykatywów (np. ‘ <i>nie wolno</i> ’)	207
4.9. Reguły inspirowane analizą korpusu diachronicznego	208
Reguła ll -> l	209
Reguła ff -> f	210
Reguła pp -> p	212
Reguła: mm -> m	213
Reguła rr -> r	214
Reguła ss -> s	216
Reguła nn -> n	218
Reguła kk -> k	219
Reguła spolszczenia wyrazów pochodzenia niemieckiego	219
Rule: “x-> gz”	220
Reguła zmodernizowania formy <i>gerundium</i>	221
Reguła zmodernizowania końcówek fleksyjnych rzeczowników	223
Reguła “a -> y” dla rzeczowników obcego pochodzenia w mianowniku liczby mnogiej	224

Reguła zmodernizowania przedrostków “vice” I “contr” . . . . .	225
Reguła zmian fonetycznych . . . . .	226
Reguła pozostałych zmian w obrębie wyrazu . . . . .	227
Reguła zmiękczenia: n->ń, s->ś . . . . .	228
Reguła zmodernizowania końcówek bezokoliczników . . . . .	229
Reguła zmodernizowania form czasowników w formie orzecznika w pierw- szej osobie liczby mnogiej . . . . .	229
Reguła zmodernizowania imiesłowu przymiotnikowego biernego . . . . .	230
4.10. Kolejność przetwarzania reguł . . . . .	230
Podsumowanie rozdziału 4. . . . .	231
5. Podsumowanie pracy . . . . .	232
6. Bibliografia . . . . .	235
7. Spis rysunków . . . . .	240