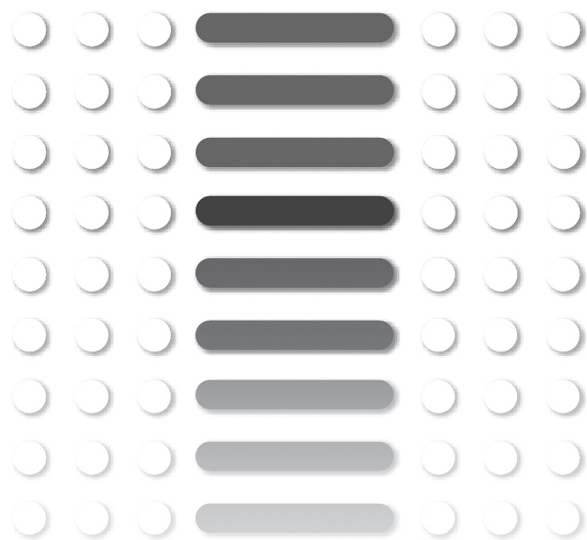


NARODOWY
KORPUS
JĘZYKA
POLSKIEGO

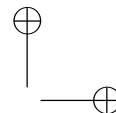
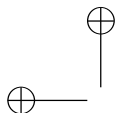
Praca zbiorowa pod redakcją
Adama Przepiórkowskiego
Mirosława Bańko
Rafała L. Górskiego
Barbary Lewandowskiej-Tomaszczyk



NARODOWY
KORPUS
JĘZYKA
POLSKIEGO



WYDAWNICTWO NAUKOWE PWN
WARSZAWA 2012



Projekt okładki i stron tytułowych
Przemysław Spiechowski

Wydawca
Magdalena Ścibor

Redaktor
Joanna Cierkońska

Produkcja
Edyta Kunowska

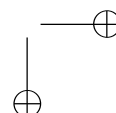
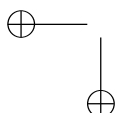


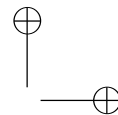
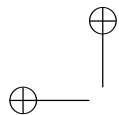
Publikacja jest dostępna na licencji Creative Commons Uznanie Autorstwa 3.0 Polska. Treść licencji dostępna jest na stronie <http://creativecommons.org/licenses/by/3.0/pl/>.

Warszawa 2012

ISBN 978-83-01-16700-4

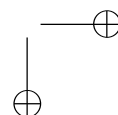
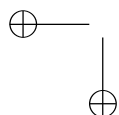
Wydawnictwo Naukowe PWN SA
02-676 Warszawa, ul. Postępu 18
tel. 22 69 54 321; faks 22 69 54 031
e-mail: pwn@pwn.com.pl; www.pwn.pl





Spis treści

I	Wstęp	1
1	NKJP: Powstanie i dzień dzisiejszy	3
1.1	Krótki rys historyczny	3
1.2	Narodowy Korpus Języka Polskiego	8
1.3	Podziękowania	9
II	Struktura korpusu	11
2	Typologia tekstów w NKJP	13
2.1	Przegląd istniejących typologii	14
2.2	Typologia tekstów w NKJP	14
2.3	Kanał	16
2.4	Klasyfikacja tematyczna	17
2.5	Definiowanie typów tekstów	18
2.6	Przypadki graniczne	21
2.7	Podsumowanie	22
2.8	Metadane	22
3	Reprezentatywność i zrównoważenie korpusu	25
4	Język mówiony w NKJP	37
4.1	Język mówiony a reprezentatywność korpusu	37
4.2	Typy języka mówionego w NKJP	38
4.3	Pozyskiwanie danych konwersacyjnych	42
4.4	Anotacja	42
4.5	Wyszukiwarka	45
4.6	Przyszłe prace	46
4.7	Podsumowanie	47



III	Zasady znakowania	49
5	Ręcznie znakowany milionowy podkorpus NKJP	51
5.1	Idea	51
5.2	Konstrukcja podkorpusu	53
5.3	Dostępność	57
6	Anotacja morfoskładniowa	59
6.1	Wstęp	59
6.2	Segmentacja	60
6.3	Tagset	62
6.4	Lematyzacja	69
6.5	Zasady znakowania	71
6.6	Anotatoria i znakowanie	81
6.7	Podsumowanie i perspektywy	95
7	Anotacja sensami słów	97
7.1	Konstruowanie słownika sensów	98
7.2	Kryteria wyodrębniania sensów „zgrubnych”	99
7.3	Sensy rzeczowników	100
7.4	Sensy czasowników	101
7.5	Sensy przymiotników	103
7.6	Wykorzystanie słownika w NKJP	104
8	Anotacja składniowa	107
8.1	Wstęp	107
8.2	Słowa składniowe	108
8.3	Grupy składniowe	114
8.4	Procedura	120
8.5	Podsumowanie	126
9	Anotacja jednostek nazewniczych	129
9.1	Nazwy własne w systemie leksykalnym polszczyzny	129
9.2	Jednostki nazewnicze w polskiej leksykografii oraz światowej i polskiej lingwistyce korpusowej	130
9.3	Modelowanie i metodologia anotacji nazw w projekcie NKJP	132
9.4	Własności jednostek nazewniczych w kontekście NKJP	145
9.5	Wnioski i perspektywy	165
10	Znakowanie XML	169
10.1	Standardy znakowania XML korpusów językowych	169

Spis treści	vii
10.2 Reprezentacja tekstu w NKJP	170
10.3 Metadane	171
10.4 Struktura tekstu	176
10.5 Segmentacja	178
10.6 Sensy słów	180
10.7 Morfoskładnia	182
10.8 Poziomy składniowe	184
10.9 Korpus ręcznie znakowany	188
10.10 Podsumowanie	193
IV Narzędzia i podprojekty	195
11 Tager morfosyntaktyczny PANTERA	197
11.1 O narzędziu	197
11.2 Algorytm	198
11.3 Adaptacja algorytmu Brilla dla języków fleksyjnych	199
11.4 Ewaluacja	203
11.5 Instrukcja obsługi	204
11.6 Wnioski końcowe	207
12 Automatyczne znakowanie sensami słów	209
12.1 Wprowadzenie	209
12.2 Opis projektu	210
12.3 Word Sense Disambiguation Development Environment	213
12.4 Przeprowadzone eksperymenty	217
12.A Dodatek. Drobne i proste klasy gramatyczne	221
12.B Dodatek. Dane ręcznie anotowanego korpusu	221
13 Narzędzia do anotacji jednostek nazewniczych	225
13.1 Anotacja wstępna metodami regułowymi – platforma SProUT	227
13.2 Ręczna poprawa anotacji podkorpusu milionowego – platforma TrEd	233
13.3 Anotacja pełnego korpusu przy użyciu uczenia maszynowego	239
13.4 Wnioski i perspektywy	252
14 Wyszukiwarka PELCRA dla danych NKJP	253
14.1 O wyszukiwarce	253
14.2 Skrócone odsyłacze	254
14.3 Składnia zapytań w przykładach	255

14.4	Sortowanie	259
14.5	Grupowanie	260
14.6	Metadane	260
14.7	Wyrazy kontekstowe	261
14.8	Analiza rejestru	262
14.9	Szeregi czasowe	262
14.10	Pobieranie wyników w postaci arkuszy kalkulacyjnych	264
14.11	Wyszukiwanie kolokacji	265
14.12	Dostęp programistyczny	270
14.13	Wyszukiwarka dla danych mówionych	272
14.14	Dalsze informacje	273
15	Słowa dnia	275
V	Zastosowania	281
16	NKJP w oczach leksykografa	283
16.1	Kartoteki w pracy nad słownikami	283
16.2	Znaczenie korpusu dla leksykografii	286
16.3	Sam korpus nie wystarczy	289
17	Zastosowanie korpusów w badaniu gramatyki	291
18	NKJP w warsztacie tłumacza	301
18.1	Rola korpusów referencyjnych	301
18.2	Ekwiwalencja frazeologiczna	302
18.3	Poprawność leksykalno-gramatyczna przekładu	306
18.4	Weryfikacja rejestru ekwiwalentu	309
18.5	Podsumowanie	310
	Bibliografia	313