Elżbieta Hajnicz

Najbardziej znane korpusy tekstów

Opracowanie przeglądowe

Nr 1021

Warszawa, grudzień 2011

Streszczenie

Niniejszy raport opisuje najbardziej znane korpusy tekstów języka naturalnego. Wpierw analizowane są zasady konstruowania korpusu, czyli doboru składających się nań tekstów w zależności od przyjetego rozmiaru oraz określenia jego struktury. Następnie prezentowane są najbardziej znane korpusy, głównie anglojęzyczne, lecz także innych języków europejskich: francuskiego, niemieckiego, rosyjskiego i czeskiego. Szczególną uwagę poświęcono dwum korpusom polskim — Korpusowi IPI PAN oraz Narodowemu Korpusowi Języka Polskiego. Oddzielny rozdział poświęcony jest bankom drzew, czyli korpusom znakowanym syntaktycznie.

Słowa kluczowe: lingwistyka komputerowa, korpusy tekstów, banki drzew, znakowanie korpusów

Abstract

Most popular text corpora. The survey

The present report describes the most famous corpora of natural language texts. First, the rules of corpora construction are analysed, namely, determining its structure and selecting texts to be included in the corpus. Next, the most popular corpora are presented. The majority of them are English corpora, but corpora of other European languages: French, German, Czech and Russian are considered as well. The special attention is paid to two Polish corpora:—the IPI PAN Corpus and the National Corpus of Polish. The separate section is devoted to treebanks, i.e., corpora that are syntactically annotated.

Keywords: computational linguistics, text corpora, treebanks, corpora annotation

Spis treści

1 Wstęp

Mianem *korpusu języka* zwykło się określać duży zbiór tekstów napisanych w jednym lub wielu językach.[1] Oczywiście nie każdy zbiór tekstów może być określony tym mianem. Nie jest nim niewątpliwie biblioteczka zgromadzona przez autorkę lub któregokolwiek z Czytelników tego tekstu. Nie są nim nawet potężne zbiory Biblioteki Narodowej zawierające wszystko, co ukazało się w języku polskim, przynajmniej w ciągu ostatnich 60-ciu lat.

Cóż więc rozumiemy pod tym terminem? Do jego wyjaśnienia powinno pomóc nam określenie, co tak naprawdę mamy na myśli używając słowa *korpus*. Władysław Kopaliński (1968) wymienia nastepujące znaczenia tego słowa: *tułów, kadłub, jednostka operacyjna wojska złożona z kilku dywizji albo brygad; korpus oficerski, dyplomatyczny, konsularny — ogół oficerów, dyplomatów czy konsulów zagranicznych akredytowanych w danym kraju.* Jak widać *korpus* jest to wyróżniona część jakiegoś zbioru (narządów, elementów konstrukcji, osób) spajająca te elementy w całość czy też determinująca ich funkcjonowanie.

Inny słownik języka polskiego Bańki (2000) definiuje *korpus tekstów* jako *zbiór książek, czasopism, artykułów itp. przeznaczony do jakichś prac lub badań.* Jest to jednak definicja bardzo ogólnikowa. Z kolei *Dictionary of the English Language* (1967) proponuje następującą definicję (za (Francis, 2007)): *zbiór wypowiedzeń bądź zdań, który winien być reprezentatywny i użyteczny przy gramatycznej analizie wybranego języka lub dialektu.* Taka definicja jest z kolei zbyt wąska. Niemniej z pełnym przekonaniem można powiedzieć, że *korpus języka* jest to podzbiór tekstów zorganizowany w taki sposób, by umożliwić sprawne wydobywanie i weryfikowanie informacji na temat dowolnego dobrze określonego podzbioru języka. W tym znaczeniu za prekursory korpusów należy uznać zbiory tzw. *fiszek*, pracowicie gromadzone przez pokolenia językoznawców przy okazji tworzenia słowników czy gramatyk. Sposób zorganizowania tych zbiorów miał służyć jak najłatwiejszemu wydobywaniu potrzebnej informacji przez badacza. Takie zbiory były jednak dobrane i uporządkowane w jeden konkretny sposób podporządkowany określonemu zadaniu językoznawczemu, i wykorzystanie ich do innego celu wymagałoby całkowitego przeorganizowania zbioru, co było zadaniem niewiele mniej pracochłonnym niż gromadzenie danych od nowa.

Termin *korpus języka* pojawił się dopiero wraz z rozwojem technik komputerowych. Chociaż początkowo komputery używane były głównie do skomplikowanych obliczeń numerycznych, wkrótce okazało się, że w wypadku wyszukiwania informacji z dużych zbiorów danych i innych operacji symbolicznych wykonywanych na dużą skalę komputer także nie ma sobie równych. Dotyczy to w szczególności informacji językowych.

[1] W niniejszym opracowaniu nie będziemy zajmować się korpusami wielojęzycznymi, choć stanowią one ważną część ogółu korpusów.

2 Budowa korpusu

Poza korpusami ogólnymi, mającymi za zadanie gromadzenie tekstów reprezentujących w możliwie wyczerpujący sposób wszystkie zjawiska występujace w języku, istnieją też korpusy specjalistyczne, gromadzące teksty dobrane według pewnej określonej reguły. Zazwyczaj są to teksty z wybranej dziedziny życia lub wiedzy. Jednak niejednokrotnie sposób doboru tekstów determinowany jest przez rodzaj planowanego zastosowania. Takim specyficznym rodzajem korpusów są korpusy diachroniczne, gromadzące wybrany rodzaj tekstów publikowanych przez zadany okres czasu, służące do śledzenia zmian następujących w języku na przestrzeni lat. Mniej typowym przykładem jest *American Heritage Intermediate Corpus* (Carroll *et al.*, 1971) będący zbiorem cytatów ważnych dla amerykańskiej kultury.

Już na etapie planowania korpusu należy odpowiedzieć na następujace pytania (Francis, 2007):

1. Jaka jest przestrzeń języka, na podstawie której korpus ma być tworzony?
2. Jaki jest planowany rozmiar korpusu?
3. Jaka ma być jego struktura?

2.1 Dobór tekstów do korpusu

W przypadku korpusów ogólnych przestrzeń, na podstawie której mamy dokonać doboru tekstów, jest niewyobrażalnie olbrzymia. I nie mówimy tu o języku mówionym. Przeciętny człowiek wypowiada w ciągu jednego dnia średnio ok. 6000–20000 słów, jednak ich ulotność nieomalże uniemożliwia ich pozyskiwanie. Korpusy języka mówionego, w tym korpusy dialogów, gromadzone są w warunkach poniekąd sztucznych, warunkowanych koniecznością ich nagrywania. Zakres większości takich korpusów jest więc ograniczony ze względu na konkretne zastosowanie. Typowy przykład stanowić tu może gromadzenie dialogów przeprowadzanych w ramach telefonicznych usług informacyjnych prowadzonych przez wiele firm (por. Marciniak, 2010); dialogi takie są bowiem standardowo nagrywane.

Jeśli jednak nawet ograniczymy nasze zainteresowanie do języka pisanego, a naszą przestrzeń do tekstów drukowanych,[2] pozostaje ona nadal olbrzymia. Całość publikacji prasowych, (dzienników, tygodników, miesięczników, wliczając w to prasę lokalną i periodyki specjalistyczne, w tym naukowe) i publikacji

[2]Dostępność notatek, pamiętników i innych rękopisów jest większa niż dostępność słowa mówionego jedynie teoretycznie.

książkowych (beletrystyki, poradników, publikacji specjalistycznych, w tym popularnonaukowych i naukowych) idzie w tysiące pozycji rocznie. A przecież nie wspominaliśmy tu o drukach akcydensowych, o dokumentach wewnętrznych firm, które też mają postać drukowaną, a więc stanowią część omawianej przestrzeni.

Należy mieć więc świadomość, że każdy korpus ogólny, niezależnie jak duży i jak starannie skonstruowany, będzie jedynie kroplą w oceanie języka. Tak więc mrzonką jest nadzieja, że za pomocą korpusu będziemy w stanie wykryć i opisać wszystkie zjawiska występujace w języku. Obowiązuje tu bowiem prawo Zipfa (1935).

Tak więc żaden korpus ogólny nie będzie zawierał wszystkich wyrazów danego języka. Istotne jest jednak, by zawierał wszystkie zjawiska (słownikowe, gramatyczne etc.) pojawiające się w języku dostatecznie często.

Oczywiście jeśli przestrzeń jest dostatecznie ograniczona, uzyskujemy pełną reprezentatywność w sposób trywialny, gromadząc w korpusie całą przestrzeń (np. całość twórczości Szekspira w badaniach Spevacka (1968–70; 1972); za (Francis, 2007)). Jednak nie stanowi to żadnej wskazówki przy konstrukcji bardziej ogólnie wyspecyfikowanych korpusów.

Rozmiar korpusu powinien być dostateczny dla reprezentacji zjawisk, których badanie jest celem jego konstrukcji. Istnieją jednak jego praktyczne ograniczenia, związane z ilością czasu i funduszy, jakie jesteśmy w stanie przeznaczyć na zgromadzenie i przetworzenie tekstów do postulowanego formatu korpusu, a nie jest tajemnicą, że jest to zadanie pracochłonne. Innym ograniczeniem, coraz mniej obecnie istotnym, są możliwości techniczne (pojemność dysków i szybkość procesorów) komputerów.

2.1.1 Zrównoważenie korpusów

Jedną z najważniejszych własności korpusów tekstów jest ich *zrównoważenie* charakteryzujące poziom reprezentatywności zgromadzonego materiału językowego w stosunku do wybranej przestrzeni językowej. Chcielibyśmy bowiem, by korpus nie tylko zawierał częste zjawiska językowe, lecz by reprezentował je proporcjonalnie do ich rzeczywistej częstości występowania, czyli poprawnie reprezentował język (lub jego wyspecyfikowany podzbiór). Tylko wówczas możliwe jest tworzenie słowników, gramatyk i innych zasobów pochodnych na bazie danego korpusu.

Tak więc po ustaleniu rozmiaru korpusu, należy wyznaczyć sposób próbkowania tekstów. Biber (2007) twierdzi, że jest to znacznie bardziej istotne niż ustalenie rozmiaru. Sugeruje nawet, że korpus powinien być tworzony przyrostowo: po ustaleniu sposobu próbkowania tworzony jest korpus próbny, po

zbadaniu jego własności zasady próbkowania są modyfikowane, i dopiero wówczas tworzony jest korpus docelowy (Biber sugeruje nawet, by dokonywać tej operacji kilkakrotnie).

Należy rozumieć, że czysto losowy dobór tekstów nie zapewnia rzeczywistego zrównoważenia korpusu. Zupełnie tak, jak w pełni losowy (a właściwie przypadkowy) dobór osób ankietowanych nie zapewnia realistycznych wyników ankiety, najpierw należy podzielić ich według płci, wieku, wykształcenia itp. Tak więc w celu zrównoważenia korpusu należy dokonać trzech operacji:

1. Podziału przestrzeni językowej na możliwie jednolite części wraz z określeniem ich procentowego udziału w korpusie;
2. Określenia rozmiaru próbek tekstów (mierzonego w liczbie wyrazów);
3. Losowania tekstów lub ich fragmentów w ramach wydzielonych podprzestrzeni.

Jedynie operacje 2 i 3 zależą od wielkości korpusu, i oczywiście są od siebie współzależne: ograniczenie rozmiaru próbki automatycznie zwiększa liczbę losowanych tekstów.[3] Operacja pierwsza zależy wyłącznie od jego charakteru, czyli własności przestrzeni językowej oraz przeznaczenia korpusu.

Aby poprawnie dokonać powyższych operacji, przynależność utworów do poszczególnych klas, na które podzielona została przestrzeń językowa, musi być dobrze i jednoznacznie zdefiniowana.

Francis (2007) zauważa, że czynnikiem rzadko branym pod uwagę przy konstrukcji korpusu jest tzw. *indeks odbioru* (ang. *reception index*) wskazujący liczbę odbiorców danego utworu. Indeks ten jest np. wyższy w wypadku gazet ogólnokrajowych niż gazet lokalnych czy specjalistycznych, więc gazety te powinny być szerzej reprezentowane. Posługiwanie się takim indeksem oznacza losowanie ważone. Jednak kierowanie się tym indeksem zbyt dosłownie wydaje się ryzykowne, zwłaszcza w wypadku korpusów ogólnych. Wybitne dzieła literackie są mniej poczytne od kryminałów i romansów, nie mówiąc już o prasie codziennej. Czy jest to wystarczający powód, by dodatkowo ograniczać ich udział w korpusie, niezależnie od ich i tak niewielkiego udziału w rynku wydawniczym?

Jeśli chodzi o rozmiar próbki, część korpusów (np. Korpus IPI PAN, por. sekcja 5) zawiera wybrane teksty w całości, w innych (np. Korpus Frekwencyjny Języka Polskiego, Kurcz *et al.*, 1990) zawarte są fragmenty ustalonej długości. Istnieje też opcja określenia górnej granicy długości tekstu; w wypadku wylosowania utworu dłuższego jest on przycinany do wymaganego rozmiaru (np. Brytyjski Korpus Narodowy; por. sekcja 3.2).

[3] Nawet wówczas, gdy dopuszczamy losowanie kilku próbek z jednego dłuższego tekstu.

Mówiąc o konstruowaniu zrównoważonego korpusu, nie można zapominać o nieuniknionych ograniczeniach praktycznych: dostępności tekstu w wersji elektronicznej[4] (tzw. „dygitalizacja" zwiększa kosztowność całej operacji) oraz bardzo istotnej, przynajmniej na polskim gruncie, kwestii praw autorskich, co wiąże się z trudnością pozyskania zgody na umieszczenie utworu w korpusie. Prowadzi to do tworzenia korpusu „oportunistycznego" składającego się ze wszystkich dostępnych utworów[5], z którego następnie losowany jest korpus zrównoważony. Korpus „oportunistyczny" może być wirtualny, przechowywanych w postaci listy utworów wraz z ich rozmiarem.

Dokładne omówienie kwestii reprezentatywności korpusów, łącznie z analizą zależności frekwencji różnych zjawisk językowych (np. części mowy lub fraz poszczególnych typów) od rodzaju tekstu, można znaleźć w (Biber, 2007).

2.2 Reprezentacja tekstów w korpusie

W wypadku korpusów ogólnych, a nawet wielu specjalistycznych, reprezentowane w nich teksty są niejednolite pod wieloma względami. Po pierwsze, mogą one być w rozmaity sposób kodowane (np. dla języka polskiego używa się współcześnie kodów cp1250, latin2, utf, jednak istnieje kilka starszych kodów, np. mazovia). Po drugie, tekst może być zapisany w różnych formatach, np. formatach edytorów (Word), systemów składu (LaTeX, Quark) czy prezentacji tekstu (PDF, HTML). Starsze teksty mogą być dostępne w formatach dawnych edytorów (np. WordStar, WordPerfect, ChiWriter). Nie wspominając o tekstach dostępnych wyłącznie w postaci drukowanej, które muszą wpierw zostać zdigitalizowane, czyli przepisane na komputerze, lub raczej zeskanowane i przetransformowane z obrazu na tekst (OCR), po czym niezbędna jest korekta błędów pisowni.

Kolejną nie mniej ważną kwestią jest podział tekstów na rozdziały, podrozdziały i akapity, sposób wersyfikacji, występowanie cytowań. Teksty mogą się w sposób istotny różnić miedzy sobą także i w tej kwestii. Edytory dostarczają też możliwość wyróżniania tekstu na różne sposoby (rozmiar i kształt czcionki, podkreślenia itp.).

Podczas wprowadzania do korpusu teksty muszą zostać ujednolicone pod wszystkimi tymi względami. Najprostszym sposobem jest transformacja do czystego tekstu w wybranej stronie kodowej. Jednak taka reprezentacja nie tylko powoduje utratę wielu informacji, lecz uniemożliwia dalsze znakowanie, bez którego uzyskany korpus byłby bardzo ubogi.

[4] Obecnie ograniczenie to jest coraz rzadsze, jednak przy dostatecznie szerokiej przestrzeni tekstów, obejmującej publikacje sprzed paru dziesiątków lat, wciąż realistyczne.

[5] Określenia tego używam za Adamem Przepiórkowskim (2004).

Dlatego korpusy reprezentowane są za pomocą bardziej wyrafinowanych schematów, głównie SGML i XML. Współcześnie zasady reprezentacji oparte są na wskazówkach tzw. *Inicjatywy Kodowania tekstów* (ang. *Text Encoding Initative*, TEI). Inicjatywa ta powstała w 1987 roku w celu opracowania i utrzymania metod hardwarowych i softwarowych służących do kodowania danych humanistycznych, w tym korpusów, w postaci elektronicznej.

Zalecany przez TEI schemat kodowania zawiera szereg modułów, z których każdy stanowi deklarację pewnego zbioru elementów XML-owych (bądź SGML-owych)[6] wraz z ich atrybutami. Schemat może być w zasadzie dowolną kombinacją modułów, jednak niektóre szczególnie ważne moduły powinny znajdować się w każdym schemacie. Moduły, które mogą znaleźć zastosowanie w reprezentacji korpusów, to header zawierający metadane, core zawierający elementy wspólne dla różnego typu dokumentów, figures służący do reprezentacji elementów nietekstowych, takich jak tabele, formuły i rysunki, namesdates zawierający jednostki identyfikacyjne oraz oczywiście corpus zawierający elementy specyficzne dla korpusów. Pierwsze dwa z tych modułów są zalecane dla dokumentów dowolnego typu. Szczegółowy opis wszystkich modułów można znaleźć w (TEI P5).

Modelowym podejściem do reprezentacji korpusów opartym na wskazaniach TEI jest *Standard Kodowania Korpusów* (ang. *Corpus Encoding Standard*, CES) (Ide, 1998b,a) sformułowany w SGML-u wraz z jego wersją XML-ową XCES (Ide *et al.*, 2000). Został on opracowany przez EAGLES (*Expert Advisory Group on Language Engineering Standards*) we współpracy z Vassar College oraz CNRS. CES został opracowany jako szeroki zbiór standardów dla wszelkich prac z zakresu przetwarzania języka naturalnego opartych na korpusach. W zamierzeniu, miał obejmować te obszary kodowania korpusów, co do których istnieje konsensus w środowisku inżynierii lingwistycznej. Tak więc CES zapewnia minimalny poziom kodowowania niezbędny do standaryzacji reprezentacji korpusów w zakresie reprezentacji opisowej tak architektury ogólnej jak i znakowania lingwistycznego. Standard taki musi być na tyle ogólny, by zminimalizować udział pracy ręcznej podczas przekształcania tekstów pochodzących z różnych źródeł do wymaganego formatu. Dlatego główną właściwością CES jest dostarczanie zestawu kodów o różnym poziomie szczegółowości, przekraczających minimalne wymagania zgodności ze standardem.

Podstawowe własności CES to:

- spójne znakowanie wszystkich tekstów korpusu na poziomie syntaktycznym i semantycznym,

[6] TEI udostępnia około 500 różnych elementów.

- możliwość przetwarzania całego korpusu za pomocą tego samego oprogramowania,
- niezależność od rodzaju zgromadzonych tekstów i przewidywanego zastosowania,
- łatwość wyszukiwania względem różnorodnych kryteriów,
- łatwość kontroli poprawności reprezentacji korpusu dostarczana przez mechanizmy SGML i XML,
- odtwarzalność (ang. *recoverability*) korpusu, czyli możność odróżnienia elementów pochodzących ze źródłowej postaci tekstu od elementów dodanych, czyli odzyskania postaci źródłowej na podstawie jej postaci zakodowanej.

Ze względu na ogólność zastosowań i wynikającą z niej elastyczność podejścia, propozycje TEI często zawierają kilka sposobów kodowania tego samego zjawiska. W ramach CES wybrano sposoby najlepiej dopasowane do kodowania korpusów i dokonano uproszczenia niektórych z nich pod kątem wspomnianych powyżej własności.

Kodowanie CES obejmuje ogólny opis podstawową strukturę tekstów, tzn. podział na tomy, rozdziały itd. aż do poziomu akapitu, wliczając w to tytuły, odnośniki, rysunki itp. oraz szczegółową strukturę tekstów obejmującą zdania, cytaty, poszczególne wyrazy, skróty, nazwy, daty itp. Z drugiej strony CES oferuje konwencję kodowania dla linwistycznego znakowania tekstów pisanych i mówionych, przede wszystkim znakowania morfosyntaktycznego oraz przyrównywanie (ang. *alignment*) korpusów równoległych. Warto zauważyć, że podstawowa struktura tekstów może być rozpoznawana niezależnie od języka, w przeciwieństwie do struktury szczegółowej.

Klasyczne podejście do reprezentacji dokumentu przeznaczonego do badań korpusowych polega na stopniowym dodawaniu kolejnych poziomów znakowania. W CES przyjęto odmienną strategię: różne poziomy znakowania przechowywane są na oddzielnych plikach SGML czy też XML (skonstruowanych zgodnie z odrębnymi plikami DTD) oraz powiązane ze sobą. Takie podejście umożliwia kilka niezależnych sposobów znakowania tego samego tekstu.

3 Najbardziej znane korpusy anglojęzyczne

Przejdę teraz do omówienia najbardziej znanych i powszechnie używanych korpusy tekstów. Dominacja języka angielskiego jest współcześnie tak duża, że nie tylko elektroniczne zasoby językowe powstawały wpierw dla języka angiel-

skiego, lecz nadal najwięcej badań z zakresu lingwistyki korpusowej (i szerzej lingwistyki komputerowej) prowadzonych jest dla tego języka. Dotyczy to także większości badań cytowanych w niniejszej pracy. Dlatego odrębna sekcja zostanie poświęcona korpusom języka angielskiego. Poza wymienionymi poniżej, w literaturze wykorzystywane są często także korpusy zgromadzone przez ACL w ramach *Inicjatywy gromadzenia danych* (ang. *Data Collection Initiative*, ACL/DCI, Liberman, 1989). Należy do nich w szczególności korpus prasowy roczników 1987–1989 Wall Street Journal, oznaczany jako korpus WSJ. Korpus zawiera podział na zdania i akapity, jednak brak mu znakowania morfosyntaktycznego.

3.1 Korpus Browna

Najstarszym elektronicznym korpusem ogólnego przeznaczenia jest Korpus Browna (ang. *Brown Corpus*). Został on stworzony pod kierunkiem W.N. Francisa i H. Kucery na Uniwersytecie Browna w Providence, RI (stąd nazwa). Zawiera on 1 mln wyrazów pisanej angielszczyzny amerykańskiej (dokładniej rzecz ujmując, prozy) wyselekcjonowanej spośród tekstów opublikowanych w Stanach Zjednoczonych w roku 1961.

3.1.1 Struktura korpusu

Korpus składa się z 500 próbek, z których każda zawiera ponad 2000 wyrazów (Francis i Kucera, 1964). Każda próbka rozpoczyna się na początku zdania, jednak nie jest to koniecznie początek akapitu, nie mówiąc już o większych jednostkach tekstu. Końcem próbki jest koniec pierwszego zdania tekstu, w którym osiągnięto 2000 wyrazów. Nie posiadam danych o najdłuższej próbce, jednak średni rozmiar próbki wynosi 2028 wyrazów. Skłania to do podejrzeń, że korpus zawiera wiele długich zdań. Mimo to można uznać, że długość próbek jest zrównoważona: nawet niezwykle długie zdanie 100-wyrazowe wydłużyłoby próbkę o ok. 4,5% powyżej średniej.

Korpus nie zawiera utworów wierszowanych, próbki beletrystyki mogą zawierać co najwyżej 50% dialogu. Tak więc w korpusie przeważają próbki tekstu ciągłego.

Selekcji próbek dokonano w dwóch etapach: wstępna subiektywna klasyfikacja tekstów wraz z określeniem, jak wiele próbek danego stylu ma zostać wyselekcjonowanych, po czym nastąpił losowy dobór poszczególnych próbek każdego stylu. Selekcji tekstów prasowych dokonano na podstawie zbiorów New York Public Library; pozostałe teksty losowano spośród zbiorów bibliotek Uniwersytetu Browna oraz Providence Athenaeum. Teksty zostały podzielone na

prozę informacyjną (374 próbki) zawierającą teksty prasowe, religijne, naukowe i popularnonaukowe oraz biografie i poradniki oraz beletrystykę (126 próbek). W sumie wyróżniono 15 podstawowych stylów językowych. Dokładniejsze dane dotyczące liczby próbek poszczególnych stylów można znaleźć w tabelach 1 dla tekstów prasowych oraz 2 dla tekstów pozostałych stylów. Ze względu na zrównoważoną długość próbek można na tej podstawie oszacować liczbę wyrazów w próbkach poszczególnych stylów, a także ich udział procentowy.

Styl tekstu	Dzienniki	Tygodniki	Razem
A. **Reportaż prasowy**			
polityczne	10	4	14
sportowe	5	2	7
społeczne	3	0	3
informacje	7	2	9
finansowe	3	1	4
kulturalne	5	2	7
Razem	33	11	44
B. **Teksty redakcyjne**			
dotycz. instytucji	7	3	10
dotycz. osób	7	3	10
listy do redakcji	5	2	7
Razem	19	8	27
C. **Recenzje prasowe**			
Razem	14	3	17

Tabela 1: Układ Korpusu Browna dla tekstów prasowych

Proporcje te mogą wydawać się zaskakujące. Dziwi zwłaszcza wysoki udział biografii (15%), dorównujący nieomal wszystkim tekstom prasowym łącznie (17,6%). Udział beletrystyki wynoszący 23,4 % można uznać za właściwy.

3.1.2 Wersje korpusu

Istnieje kilka wersji korpusu Browna. Wersja A to oryginalna wersja korpusu przygotowana w latach 1963–64. Została ona zakodowana zgodnie z procedurą kodowania opisaną w (Newman *et al.*, 1959). Kodowanie to dzieli tekst na 80-znakowe linie, z których pierwsze 70 znaków to faktyczny tekst (przekodowany na wielkie litery), 71. kolumna jest pusta, zaś ostatnie 9 znaków to kod linii.

Styl tekstu	l. próbek	Styl tekstu	l. próbek
I. Proza informacyjna		**II. Beletrystyka**	
D. **Religia**		K. **Literatura piękna**	
książki	7	powieści	20
periodyki	6	opowiadania	9
traktaty	4	**Razem**	29
Razem	17		
E. **Poradniki**		L. **Kryminały**	
książki	2	powieści	20
periodyki	34	opowiadania	4
Razem	36	**Razem**	24
F. **Popularnonaukowe**		M. **Science fiction**	
książki	23	powieści	3
periodyki	25	opowiadania	3
Razem	48	**Razem**	6
G. **Biograficzne itp.**		N. **Przygodowe**	
książki	38	powieści	15
periodyki	37	opowiadania	14
Razem	75	**Razem**	29
H. **Różne**		P. **Romanse**	
Dokumenty rządowe	24	powieści	14
Raporty fundacji	2	opowiadania	15
Raporty firm	2	**Razem**	29
Katalog uczelni	1		
Statut firmy	1		
Razem	30		
J. **Naukowe**		R. **Humorystyczne**	
Przyrodnicze	12	powieści	3
Medyczne	5	eseje itp.	6
Matematyczne	4	**Razem**	9
Społeczne	14		
Prawo i n. polityczne	15		
Humanistyczne	18		
Techniczne	12		
Razem	30		

Tabela 2: Układ Korpusu Browna dla pozostałych tekstów

Pierwsze 4 znaki to numer linii w próbce, kolejne dwie to 'E1' (pierwszy korpus angielszczyzny), zaś ostatnie 3 to kod próbki złożony z litery kodującej styl (zgodnie z oznaczeniami z tabel 1 i 2) oraz dwucyfrowy numer próbki w stylu.

Już w tej pierwotnej wersji istnieją pewne specjalne oznaczenia, dotyczące głównie struktury tekstu. Są one kodowane za pomocą dwóch gwiazdek, po których następuje znak. I tak, '**.' oznacza kropkę kończąca skrót ('**..' na końcu zdania), formuły zastępowane są przez '**F'; w podobny sposób kodowane są różne znaki specjalne (w tym przestankowe). Liczby rzymskie przekodowywane są na arabskie i objęte nawiasami z symboli '*/' i '*,' (np. '*/8*' oznacza VIII).

ozn.	opis	ozn.	opis
**N	początek tytułu	**=	początek pogrubienia
**P	koniec tytułu	**$	koniec pogrubienia
**R	początek podtytułu	**(	początek kapitalizacji
**T	koniec podtytułu	**)	koniec kapitalizacji
*=	początek kursywy	**Q	początek cytowania
*$	koniec kursywy	**U	koniec cytowania

Tabela 3: Kodowanie struktury tekstu w Korpusie Browna

Analogiczne znakowanie stosowane jest do określania struktury tekstu (por. tabela 3). Tytuł oznaczany jest dla fragmentów najwyższego poziomu (np. rozdziałów) w danej próbce, podtytuł dla wszystkich fragmentów rzędu niższego. Brak tytułu fragmentu danego poziomu oznaczany jest jako pusty (**N**P bądź **R**T, odpowiednio). Taki sposób kodowania uniemożliwia oznaczanie zagnieżdżeń.

Wersja B, utworzona w 1967 roku, została „oczyszczona" z większości znaków specjalnych i przestankowych. Miało to ułatwiać badania dotyczące pojedynczych słów, np. (Kucera i Francis, 1967). Istnieją też dwie wersje opracowane na uniwersytecie w Bergen. Istotną wprowadzoną tam zmianą jest rozróżnienie małych i wielkich liter. Kolejna wersja powstała na Uniwersytecie Stanford. Złożona jest ona z rekordów różnej długości, z których każdy odpowiada pojedynczemu zdaniu.

Najciekawszą z naszego punktu widzenia wersją jest jednak wersja C, gdyż jest ona oznakowana za pomocą zestawu 81 znaczników, z których każdy określa pewną klasę wyrazów. Istnieje 6 rodzajów klas wraz z odpowiednimi znacznikami:

a) klasy podstawowe, zwane *otwartymi klasami leksykalnymi*: rzeczowniki (pospolite `NN` i własne `NP`), czasowniki (zwykłe `VB`, modalne `MD`, być (ang.

to be) BE, mieć (ang. *to have*) HV i robić (ang. *to do*) DO), przymiotniki (zwykłe JJ, semantycznie w stopniu najwyższym (np. *główny, zasadniczy*) JJS), przysłówki (zwykłe RB, nominalne (np. ang. *here, then*) RN), partykuły (np. ang. *off, out, up*) RP, kwalifikatory (np. ang. *rather, very*) QL, prekwalifikatory (np. ang. *quite, rather*) ABL i postkwalifikatory (np. ang. *enough, indeed*) QLP oraz liczebniki (główne CD i porządkowe OD);

b) wyrazy funkcyjne (zwane *zamkniętymi klasami leksykalnymi i gramatycznymi*): określniki (rodzajniki (*a/the*) AT), określniki deiktyczne (np. ang. *this, another*) DT, określniki podwójne (ang. *either, neither*) DTX, kwantyfikatory (np. ang. *some, any*) DTI, przedkwantyfikatory (ang. *prequantifiers*, np. *all, half* ABN, w tym wyróżniony przedkwantyfikator podwójny *both* ABX i postkwalifikatory (np. ang. *many, next, single*) AP), przyimki IN spójniki (współrzędne (np. ang. *and, or, either*) CC i podrzędne (np. ang. *as, whether, like*) CS), zaimki (osobowe w l.poj. os.3 (ang. *he, she, it*) PPS, osobowe pozostałe (np. ang. *I, they*) PPSS, dzierżawcze określniki (np. ang. *my, our* PP$) i zaimki (np. ang. *mine, ours*) PP$$, zaimki zwrotne (np. ang. *itself*) PPL, zaimki na pozycji biernikowej (np. ang. *me, them*), względne zaimki (np. ang. *who, what, that*) WP, określniki (np. ang. *which, what*) WDT, kwalifikatory (np. ang. *how, however*) WQL i przysłówki (np. ang. *WRB*) oraz inne zaimki (np. ang. *one, something*) PN) oraz wykrzykniki (np. ang. *bang, oops, goodbye*) UH;

c) wybrane ważne wyrazy;

d) znaki interpunkcyjne (z których większość posiada odrębny znacznik);

e) morfemy fleksyjne: liczba mnoga S i dzierżawczość $ rzeczowników, forma biernikowa zaimków O, stopień wyższy R i najwyższy T przymiotników i przysłówków oraz czas przeszły D, 3 os. l.poj. Z, imiesłów przeszły N oraz współczesny (gerundium) G czasowników;

f) dwa dodatkowe znaczniki oznaczające wyrazy obce FW bądź cytowane NC, które mogą być łączone z innymi znacznikami determinującymi chatrakter tych wyrazów.

Oznakowanie korpusu miało na celu ułatwienie jego automatycznej analizy syntaktycznej, i sposób znakowania został podporządkowany temu celowi. Znaczniki morfemów fleksyjnych nie funkcjonują samodzielnie, lecz są zawsze dołączane do podstawowych znaczników, które zdaniem twórców systemu znakowania tworzą odrębne klasy gramatyczne. Znaczniki mogą być też łączone na inne sposoby wykorzystywane przy opisie zapisów skróconych wyrazów łączonych za pomocą apostrofu (np. ang. *I've*).

Powyższy system, niezmiernie szczegółowy i bogaty, nie jest niestety dostatecznie klarowny. Zwłaszcza klasyfikacja i znakowanie zaimków jest mało czytelne: zaimki osobowe i względne nie występują w swoim „podstawowym" kodowaniu `PP, WP`, wręcz przeciwnie, dla oznaczenia mianownikowej formy zaimka w 3. os. l.poj. używa się oznaczenia `PPS` (oznaczenie `WPS` obejmuje l.poj. i mn.), choć `S` zazwyczaj oznacza l.mn.

Podzbiór korpusu Browna zawierający ok. 250 tys. wyrazów został przez twórców wordnetu prinstońskiego (Miller *et al.*, 1990; Fellbaum, 1998) ręcznie oznakowany sensami wordnetowymi. Korpus ten (Miller *et al.*, 1993) znany jest w literaturze pod mianem SemCor.

Jeśli chodzi o angielszczyznę brytyjską, istnieje stworzony niewiele później i obejmujący publikacje z tegoż 1961 roku korpus Lancaster-Oslo/Bergen (ang. *Lancaster-Oslo/Bergen Corpus of British English*, LOB), por. (Johansson *et al.*, 1978).

3.2 Narodowy Korpus Brytyjski

Najbardziej chyba znanym i najczęściej cytowanym w literaturze jest Narodowy Korpus Brytyjski (ang. *British National Corpus*, BNC). Jego rola jest tak istotna, że nie będzie wielką przesadą uznać go za korpus wzorcowy. BNC jest omawiany m.in. w (Aston i Burnard, 1998; Burnard, 2007).

Narodowy Korpus Brytyjski jest korpusem współczesnej angielszczyzny, którego pierwsza podstawowa wersja tworzona była w latach 1990–1994 przez konsorcjum złożone ze znanych wydawnictw Oxford University Press, Longman i Chambers oraz ośrodków naukowych uniwersytetów w Oksfordzie i Lancaster i Biblioteki Brytyjskiej. W 1998 została publicznie udostępniona tzw. „światowa" wersja korpusu (ang. *the BNC World Edition*).

3.2.1 Struktura korpusu

BNC gromadzi 4054 utwory (mówione i pisane) podzielone na ok. 5 mln zdań[7] zawierających 100,467 tys. wyrazów, w tym 97,620 tys. tzw. *w-units*, czyli wyrazów oznakowanych kategoriami gramatycznymi (ang. *POS-tagging*) za pomocą tagera CLAWS (Garside, 1996). Przyjęto założenie, że długość próbki pojedynczego utworu nie może przekraczać 45 tys. słów. Jak widać, próbki mogą być

[7] Właściwie są to tzw. *s-units*, gdyż granice zdań zostały w tekście wyznaczone automatycznie i nie do końca są wiarygodne.

o rząd większe niż próbki w Korpusie Browna. Brak dolnego ograniczenia powoduje jednak, że ich rozmiar może być niezmiernie zróżnicowany. Wydaje się jednak, że takie zrównoważenie ma sens jedynie przy niewielkim przyjętym rozmiarze próbki, w przeciwnym razie zbyt wiele tekstów (np. prasowych) byłoby zbyt krótkich i nie spełniałoby kryteriów doboru.

Przyjęty rozmiar próbki powoduje, że przy stukrotnie większej liczbie wyrazów, liczba utworów w BNC jest jedynie 8 razy większa niż w korpusie Browna.

Tabela 4 ukazuje podział tekstów według sposobu ich publikacji. W przypadku języka mówionego określenie *demograficzny* oznacza zbiór nieformalnych rozmów przeprowadzonych przez wolontariuszy zrównoważony ze względu na płeć, płeć, klasę społeczną i miejsce zamieszkania. Z kolei *zależny od okoliczności* (ang. *context-govern*) oznacza dialogi zebrane w określonych okolicznościach (np. oficjalne spotkania polityczne i biznesowe, wywiady, konferencje). *Przemówienia* są to teksty pisane w celu ich późniejszego wygłoszenia. Liczba wyrazów (*w-units*) podawana jest w milionach.

rodzaj tekstu	teksty	w-units	s-units	proc.
demograficzny	153	4,30	610 593	10,08
zależny od okoliczności	757	6,28	428 558	7,07
mówiony razem	910	10,58	1039 121	17,78
książki i periodyki	2688	80,49	4403 803	72,75
przemówienia	35	1,35	120 153	1,98
pisany różne	421	7,55	490 016	8,09
pisany razem	3144	89,39	5013 972	82,82

Tabela 4: Układ Brytyjskiego Korpusu Narodowego

Tabela 5 przedstawia podział części pisanej korpusu na dziedziny. Podział ten jest poniekąd ortogonalny do zaprezentowanego w poprzedniej tabeli. Obie tabele pochodzą z pracy (Burnard, 2007). Dziwić może jedynie, że w drugiej z nich nie została wyróżniona problematyka krajowa, której przecież nie można utożsamiać z gospodarką. Określenia *beletrystyka* i *przemyślenia* są moimi nieręcznymi i upraszczającymi tłumaczeniami terminów *imaginative* oraz *belief and thought*.

W przeciwieństwie do korpusu Browna, BNC zawiera transkrypcje tekstów mówionych (17,8%). Także jeśli chodzi o teksty pisane, trudno jest porównywać te dwa korpusy ze względu na trochę inny zakres (np. w korpusie Browna brak przemówień) oraz odmienny sposób prezentacji danych. Pomimo mniejszego

dziedzina	teksty	w-units	proc.	s-units	proc.
beletrystyka	477	16 377 726	18.76	1356 458	27.05
gospodarka i finanse	295	7 257 542	8.31	382 717	7.63
nauki matematyczno-przyrodnicze	146	3 784 273	4.33	183 466	3.65
nauki społeczne	527	13 906 182	15,93	700 122	13.96
nauki stosowane	370	7 104 635	8,14	357 067	7.12
przemyślenia	146	3 007 244	3.44	151 418	3.01
rekreacja	438	12 187 946	13.96	760 722	15.17
sprawy międzynarodowe	484	17 132 023	19.62	800 560	8.09
sztuka	261	6 520 634	7,47	321 442	6.41

Tabela 5: Podział Brytyjskiego Korpusu Narodowego na dziedziny

rozmiaru korpusu Browna, jego struktura ukazana w (Francis i Kucera, 1964) jest znacznie bardziej dokładna. W szczególności, Burnard (2007) nie wyróżnił tekstów prasowych. Na podstawie posiadanych danych mogę jedynie ocenić, że udział beletrystyki w obu korpusach jest zbliżony, natomiast BNC zawiera nieco mniej tekstów naukowych (24,7%).

Zgodnie z wskazówkami przyjętymi przez TEI, z każdym plikiem tekstowym korpusu powiązany jest plik nagłówkowy (ang. *header*), w którym zostały zebrane wszelkie dotyczące. Taki plik zawiera tzw. *metadane*, czyli informacje identyfikujące i klasyfikujące tekst (autor, tytuł, rok wydania itp.) wraz z danymi pomocniczymi, takimi jak dane demograficzne rozmówców dla tekstów mówionych, rozmiar, wprowadzane poprawki itp. Stworzony został także plik nagłówkowy dla całego korpusu, zawierający m.in. jego dokumentację itp.

3.2.2 Znakowanie korpusu

Jak zostało wspomniane na początku, BNC został oznakowany automatycznie przy użyciu tagera CLAWS4, którego pierwotna wersja CLAWS1 została opracowana do oznakowania korpusu LOB.

Ponadto podczas pilotażowego etapu konstrukcji korpusu 220 tys. wyrazów zostało oznakowane semantycznie na podstawie słownika Hector (Atkins, 1991).

3.3 Narodowy Korpus Amerykański

Po 30-tu latach od powstania, korpus Browna przestał wystarczać do reprezentacji angielszczyzny amerykańskiej. Z drugiej strony BNC nie mógł spełniać tej roli ze względu na zbyt wielkie różnice pomiędzy angielszczyzną amerykań-

ską a brytyjską. Dlatego w drugiej połowie lat 90-tych zaczęła się formować koncepcja utworzenia nowego, dużego korpusu angielszczyzny amerykańskiej,[8] który pod względem rozmiaru i struktury ma być zbliżony do BNC (Fillmore *et al.*, 1998). Narodowy Korpus Amerykański (ang. *American National Corpus*, ANC) ma być duży (100 mln. wyrazów), współczesny (teksty opublikowane po 1990 roku), zróżnicowany pod względem stylów językowych oraz jednolicie oznakowany pod względem wielu zjawisk językowych. Poza 100 milionową częścią statyczną, która ma pozostać niezmienna, planowane jest opracowanie części dynamicznej, złożonej z utworów dokładanych w regularnych odstępach czasu. Planowane jest uzupełnianie korpusu o 10% tekstów (w stosunku do wspomnianego powyżej rozmiaru podstawowego) różnorodnych pod względem stylu i dziedziny w odstępach pięcioletnich, co doprowadzi do powstania korpusu, który może być traktowany zarówno jak synchroniczny jak i diachroniczny. Twórcy korpusu zamierzają tu uwzględnić m.in. bogate zasoby dostępne przez internet. Teksty internetowe różnią się od tych udostępnianych za pośrednictwem tradycyjnych mediów pod względem słownictwa, skrótów itp., a nawet konstrukcji gramatycznych, i uwzględnienie tego stylu w dużym korpusie ułatwi badania językowe tego medium.

3.3.1 Wersje korpusu

ANC jest wciąż korpusem *in statu nascendi*. Jego pierwsza wersja udostępniona w 2003 roku zawiera ok. 11 mln. wyrazów (8 mln. tekstów pisanych i 3 mln. transkrypcji dialogów) (Ide i Suderman, 2004; Reppen i Ide, 2004). Składa się ona z tekstów uzyskanych przez konsorcjum na początku, w rezultacie nie jest to korpus zrównoważony. Jego struktura została utworzona automatycznie bez ręcznej walidacji. Korpus udostępniany jest w dwóch postaciach: w pierwszej każdy poziom znakowania przechowywany jest oddzielnie (zgodnie z zaleceniami CES), druga stanowi połączenie wszystkich poziomów znakowania na jednym pliku.

Druga i jak na razie ostatnia wersja ANC udostępniona w 2008 roku zawiera już ponad 22 mln. wyrazów. Jej struktura przedstawiona jest w tabeli 6 na podstawie strony korpusu `www.americannationalcorpus.org`. Należy zwrócić uwagę, że podana w tabeli liczba plików może być większa od faktycznej liczby utworów uwzględnionych w korpusie, wydaje się bowiem, że niektóre z nich zostały umieszczone w całości, a niektóre we fragmentach. Tabela 6 może sugerować duże zróżnicowanie korpusu; niestety utwory z poszczególnych dziedzin są dosyć jednorodne. Na przykład wszystkie teksty gazetowe pochodzą z wydań New York Timesa z nieparzystych dni lipca 2002.

[8] Przyjęto założenie, że ANC może zawierać wyłącznie utwory autorstwa osób urodzonych bądź wykształconych a zarazem żyjących obecnie w USA (Ide *et al.*, 2002).

Dziedzina	l. plików	l. wyrazów
Teksty pisane		
dokumenty rządowe	17	281 093
przewodniki	179	1 012 496
biomedyczne	1089	3 758 884
blogi	143	3 093 075
beletrystyka	107	466 941
korespondencja	245	91 318
gazety	4148	3 625 687
czasopisma	4563	4 821 192
informacyjne	45	330 524
Razem pisane	10821	18 530 122
Teksty mówione		
telefoniczne	2331	3 072 009
twarzą w twarz	93	198 295
dyskusje akademickie	50	593 288
Razem mówione	2474	3 863 592
Razem	13295	22 393 704

Tabela 6: Struktura Narodowego Korpusu Amerykańskiego

Stopniowe udostępnianie korpusu umożliwi jego udoskonalanie na podstawie informacji zwrotnych uzyskiwanych przez użytkowników. W przyszłości planowane jest też stworzenie podkorpusu obejmującego 10% ANC (10 mln. wyrazów), który ma być ręcznie poprawiony pod każdym względem (metadane, segmentacja, znakowanie morfosyntaktyczne).

3.3.2 Kodowanie i znakowanie korpusu

ANC jest kodowany w XML-u zgodnie z standardem XCES (por. sekcja 2.2). Pliki nagłówkowe są minimalne, zawierają jednak pełną informację dotyczącą dziedziny, poddziedziny, tematyki, odbiorców i środka przekazu. Segmentacja dokonana została do poziomu zdań. Natomiast znakowanie morfosyntaktyczne zrealizowane jest na kilka sposobów: za pomocą tagsetu POS opracowanego przez Bibera 1988; 1995 (i to znakowanie zostało uznane za podstawowe, i jako takie występuje połączonej postaci korpusu), za pomocą użytego do znakowania BNC tagera CLAWS (z tym samym tagsetem),[9], za pomocą pakietu GATE,

[9]Uzyskana w ten sposób wersja jest w szczególności przeznaczona do badań porównawczych z BNC.

który posługuje się tagsetem wykorzystanym do znakowania banku drzew Penn (ang. *Penn TreeBank*) oraz za pomocą tagera Multext opartego na tagsecie wprowadzonym w ramach projektu EAGLES. Znakowanie uzyskane za pomocą każdego tagera przechowywane jest na oddzielnym pliku co umożliwia ich niezależne wykorzystywanie.

Twórcy ANC planują w ramach rozwoju ANC dalsze poziomy znakowania, w tym wyznaczania jednostek identyfikacyjnych i relacji anaforycznych, znakowania syntaktycznego i semantycznego (w tym powiązania z kategoriami występującymi w wordnecie prinstońskim (Fellbaum, 1998) i FrameNecie (Baker *et al.*, 1998, 2003; Fillmore *et al.*, 2001, 2003).

4 Korpusy innych języków europejskich

Niezależnie od wspomnianej powyżej dominacji angielszczyzny, bogactwo zasobów stworzonych do dnia dzisiejszego dla wielu innych języków, w tym europejskich, jest imponujące.

4.1 Korpusy niemieckojęzyczne

Liczba korpusów niemieckojęzycznych jest na tyle duża, że trudno jest wymienić je wszystkie.

Ciekawym przykładem jest stworzony w Instytucie Języka Niemieckiego (*Institut für Deutsche Sprache*) w ramach projektu COSMAS II[10] (ang. *Corpus Search, Management and Analysis System*) zestaw (zwany Manheimskim — *Manheimer Corpus collection*) 89 korpusów zawierających łącznie 3.8 mld. wyrazów. Korpusy te są jednorodne tematycznie, np. zawierają teksty z konkretnego czasopisma, literaturę biograficzną, zbiór utworów Goethego, beletrystykę lat 20-tych XX w. itp. Korpusy są jednolicie kodowane w formacie XCES, jednak jedynie 3 z nich (LIMAS; Mannheimer Morgen 1991, 1994–1996; Der Spiegel, Heft 1/93 – Heft 52/94) są znakowane morfosyntaktycznie.

Najważniejszym korpusem języka niemieckiego jest korpus opracowany w ramach projektu DWDS (niem. *Das Digitales Wörterbuch der deutschen Sprache*, Ćavar *et al.* (2000))[11], uznawany za korpus narodowy. Korpus obejmuje teksty XX-wieczne i jest podzielony na dwie części, podkorpus zrównoważony złożony z 79 830 dokumentów zawierających 100 mln. wyrazów, oraz znacznie większy, zawierający ok. 500 mln. wyrazów, podkorpus oportunistyczny. Struktura korpusu zrównoważonego przedstawiona w tabeli 7 ma przypominać strukturę BNC (sekcja 3.2).

[10] `http://www.ids-mannheim.de/cosmas2/`
[11] `http://www.dwds.de/textbasis`

styl tekstu	udział	styl tekstu	udział
beletrystyka	25%	teksty luźne	20%
teksty prasowe	25%		
teksty naukowe	20%	teksty mówione	10%

Tabela 7: Układ narodowego korpusu języka niemieckiego

Pozyskane teksty są przechowywane są w formacie XML/TEI. Pliki XML zawierają metadana oraz segmentację tekstu. Znakowanie morfosyntaktyczne oraz płytkie znakowanie syntaktyczne (frazy rzeczownikowe i przyimkowe) przeprowadzone zostało za pomocą tagera MPRO (Maas, 1996, 1998; Maas *et al.*, 2009). Ponadto zostało przeprowadzone znakowanie semantyczne na poziomie wyrazowym.

Istnieje też Szwajcarski Korpus Tekstów (*Schweizer Text Korpus*, CHTK), zawierający 20 mln. wyrazów współczesnej niemczyzny szwajcarskiej (Bickel *et al.*, 2009). Jest on pomyślany jako składnik korpusu równoległego dla współczesnych tekstów niemieckich pochodzących poza tym z Niemiec, Austrii i Włoch (Tyrol Południowy). Korpus zawiera teksty z wielu dziedzin (szczegółowa lista wymienia ich 36), w tym teksty prasowe, literackie i naukowe.

4.2 Korpusy francuskojęzyczne

W ramach projektu prowadzonego w zespole Uniwersytetu Stendhala z Grenoble oraz centrum badawczego Xerox-a XRCE powstał korpus zawierający ponad 1 mln wyrazów. Korpus zawiera

- 2 książki (202 581 wyrazy) i 204 artykuły (201 280 wyrazów) z zakresu nauk humanistycznych;
- 14 artykułów popularno-naukowych z *CNRS Info* (111 886 wyrazy);
- 136 artykułów ekonomicznych z *Le Monde* (180 760 wyrazów);
- 13 numerów *Official Journal of the European Communities* (337 000 wyrazów)

Jak widać, korpus nie zawiera beletrystyki.

Korpus został podzielony na rozdziały, akapity i zdania. Segmentacja zdań dokonana została automatycznie. Korpus jest znakowany zależnościami anaforycznymi (w formacie XML).

4.3 Korpusy języka rosyjskiego

4.3.1 Korpus referencyjny

Znanym korpusem języka rosyjskiego jest *Korpus referencyjny języka rosyjskiego* (BOKR, Sharoff, 2004). Zawiera 1 mln wyrazów współczesnego języka rosyjskiego, o strukturze z założenia zbliżonej do BNC (patrz tabela 8).

styl tekstu	udział	styl tekstu	udział
życie	30%	nauki przyrodnicze	5%
polityka	15%	gospodarka	5%
nauki społeczne	12%	sztuka	5%
nauki stosowane	10%	religia i filozofia	3%
teksty luźne	10%	teksty mówione	5%

Tabela 8: Układ korpusu referencyjnego języka rosyjskiego

Korpus ten jest reprezentowany w formacie SGML zgodnym z TEI. Jest on znakowany morfosyntaktycznie. Zastosowany tagset jest oparty na tagsecie MULTEXT-East (Erjavec, 2001; Erjavec *et al.*, 2003). Zgodnie z nim, każda klasa gramatyczna (wyróżnione jest 14 klas) posiada odrębny zestaw znaczników (kategorii gramatycznych), zestawy te nie są jednak rozłączne dla różnych klas gramatycznych. Wartości znaczników rozważonych w MULTEXT-East uwzględniają specyfikę wszystkich języków, dla których tagset ten był tworzony (słoweński, czeski, bułgarski, rumuński, estoński i węgierski; rosyjski wraz z chorwackim i serbskim zostały dołączone później). W ten sposób np. kategoria przypadek może mieć przypisanych 30 różnych wartości! Oczywiście tagsety poszczególnych języków ograniczane są do wartosci, które w nich występują, jednak sposób kodowania jest wspólny.

4.3.2 Korpus narodowy

Jak wskazuje nazwa, najważniejszym korpusem języka rosyjskiego jest *Narodowy Korpus Języka Rosyjskiego* (ros. *Narodnyj Korpus Russkogo Jazyka*, NKRJ, (Grishina i Rakhilina, 2005), `www.ruscorpora.ru`). Powstał on w Instytucie Języka Rosyjskiego Rosyjskiej Akademii Nauk. Korpus gromadzi ok. 52 tys. utworów zawierających łącznie ok. 150 mln. segmentów. Utwory opatrzone są szczegółowymi metadanymi, np. dane o autorze zawierają poza imieniem i nazwiskiem także datę urodzenia, płeć,[12] i wiek w momencie napisania utworu.

[12]Informacja redundantna względem imienia i nazwiska, jednak potencjalnie użyteczna przy manipulowaniu zasobami korpusu.

Podstawę korpusu reprezentujący teksty standardowego języka korpusu można podzielić na 3 części:

- współczesny język pisany zawierający teksty powstałe po 1950 roku,
- dawny język pisany zawierający teksty od połowy XVIII do połowy XX wieku,
- współczesny język potoczny zawierający zapisy tekstów mówionych, oficjalnych i spontanicznych.

W NKRJ zostały wyróżnione dwa podkorpusy:

- korpus zrównoważony zawierający 150 mln. segmentów
- korpus znakowany ręcznie zawierający 5 mln. segmentów.

styl tekstu	udział	styl tekstu	udział
beletrystyka	40%	**niefikcyjne**	60%
autobiografie	1%	elektroniczne	2%
dramat	2%	filozoficzne	3%
humorystyczne	4%	gospodarcze	3%
fantastyka	8%	naukowe	22%
kryminały	9%	prasowe	68%
lit. dziecięca	5%	reklamowe	1%
proza historyczna	6%	techniczne	1%
przygodowe	3%	życie codzienne	2%
romanse	2%		
inne	60%		

Tabela 9: Struktura Narodowego Korpusu Języka Rosyjskiego

W tabeli 9 widnieje struktura Narodowego Korpusu Języka Rosyjskiego. Jest on podzielony na dwie cześci, beletrystyczną obejmującą 40% materiału językowego i pozostałą obejmującą 60% objętości korpusu. Udział poszczególne podstylów wyliczany jest względem tych wartości, co oznacza w szczególności, że teksty prasowe zawierają 41% materiału językowego, tyle ile beletrystyka.

Znakowanie morfosyntaktyczne Znakowanie morfosyntaktyczne całego korpusu zrealizowane zostało automatycznie. Podstawę znakowania stanowi (Zalizniak, 1977). Znaczniki morfosyntaktyczne przypisywane formie wyrazowej dzielą się na cztery grupy. Pierwsza zawiera lemat i klasę gramatyczną

wyrazu, na drugą składają się cechy gramatyczne leksemu (takie jak rodzaj rzeczownika), trzecia to cechy morfoskładniowe danej formy (np. przypadek), wreszcie w czwartej umieszczane są informacje dotyczące niestandardowych cech wyrazu, np. wariant ortograficzny.

Semantyka Większość wyrazów w podkorpusie zrównoważonym (należących do otwartych klas gramatycznych) jest oznaczona semantycznie za pomocą znaczników semantycznych i derywacyjnych (Apresjan *et al.*, 2006; Lashevskaja, 2006; Kustova *et al.*, 2007). Znakowanie zostało zrealizowane za pomocą programu `Semmarkup` autorstwa A. E. Poliakova wykorzystującego słownik semantyczny korpusu. Program nie rozstrzyga wieloznaczności.

Słownik semantyczny oparty został na bazie leksykograficznej *Lexicograph* (Filipenko *et al.*, 1992, `www.lexicograph.ru`) pod kierunkiem E. Paduchevej i E. Rakhiliny, jednak zbiór rozważanych klas semantycznych został istotnie rozszerzony. Użyte oznaczenia oparte zostały na notacji angielskiej.

Zestaw znaczników zależy od klasy gramtycznej. Rzeczowniki zostały podzielone na trzy grupy: konkretne (`r:concr`), abstrakcyjne (`r:abstr`) i nazwy własne (`r:propn`). Analogiczny podział rozważa Szupryczyńska (1973), tyle że nie rozważa ona nazw własnych. Przymiotniki podzielone zostały na cztery grupy: jakościowe (`r:qual`), relacyjne (`r:rel`), dzierżawcze (`r:pos`) i nieodmienne (`r:inv`).

Znaczniki są podzielone w następujący sposób:

- taksonomiczne: dla rzeczowników „osoba" (`t:hum`), „substancja" (`t:stuff`), „przestrzeń" (`t:space`), „ruch" (`t:move`) itp.; dla czasowników „ruch" (`t:move`), „położenie" (`t:loc`), „percepcja" (`t:perc`) itp.; dla przymiotników i przysłówków „prędkość" (`t:speed`), „ rozmiar" (`t:size`), „położenie" (`t:place`);
- mereologiczne: „część" (`pt:part`), „zbiór" (`pt:set`), dla rzeczowników konkretnych i abstrakcyjnych;
- topologiczne: „pojemnik" (`top:contain`), „powierzchnia pozioma" (`top:horiz`), dla rzeczowników konkretnych;
- przyczynowe dla czasowników;
- derywacyjne: „zdrobnienie" (`d:dim`), „zgrubienie" (`d:aug`), „przymiotnik odczasownikowy" (`der:v`);
- ewaluacyjne: „pozytywne" (`ev:posit`) i „negatywne" (`ev:neg`), dla rzeczowników konkretnych i abstrakcyjnych oraz przymiotników i przysłówków.

Rodzaje znaczników do pewnego stopnia przypominają zestaw relacji wordnetowych — hiponimię, meronimię, derywację — tyle że w przeciwieństwie do wordnetu nie mamy tu do czynienia z relacjami.[13]

Znaczniki te mogą mieć konstrukcję hierarchiczną, np. `t:hum:etn` dla narodowości, `t:tool:mus` dla instrumentów muzycznych (rzeczowniki), `t:time:dur:max` — długotrwałość (przymiotniki i przysłówki), `t:impact:creat` — tworzenie obiektów fizycznych, `t:psych:emot` — emocje, itp. Klasyfikacja taksonomiczna jest w dużym stopniu zbliżona do kategorii semantycznych *Słowosieci*, tyle że kategorie te nie miały struktury hierarchicznej.

Bank drzew Fragment Narodowego Korpusu Języka Rosyjskiego posłużył do stworzenia banku drzew (Boguslavsky *et al.*, 2000, 2002; Nivre *et al.*, 2008) SynTagRus zawierającego około 30 tys. zdań współczesnego języka rosyjskiego zróżnicowanego pod względem stylu (10% beletrystyka, 78% teksty prasowe, 12% wiadomości internetowe). Składające się nań zdania tworzą drzewa rozbiorów zależnościowych, w których węzły stanowią wyrazy składające się na zdanie, zaś krawędzie etykietowane są nazwami relacji syntaktycznych.

Struktura składniowa banku oparta jest na Modelu Sens $\Longleftrightarrow$ Tekst (ang. *Meaning⇔Text Theory*) zaproponowanym przez Igora Mel'čuka i Alexandra Zholkovsky'ego (Mel'čuk i Zholkovsky, 1984; Mel'čuk, 1988). Połowa z ok. 80 relacji zależnościowych została zaproponowana w powyższym modelu. Relacje dzielą się na 6 grup: czynnościowe (łączą predykat z argumentami), atrybutywne (zazw. łączą rzeczownik z modyfikatorem), ilościowe (łączą rzeczownik z kwantyfikatorem, np. liczebnikiem), przysłówkowe (łączą predykat z modyfikatorem przysłówkowym), koordynacyjne (łączą frazy i zdania koordynowane przez spójniki) oraz pomocnicze (zazw. łączą dwa elementy tworzące jedną jednostkę syntaktyczną, np. analityczną postać czasownika). Struktura zdania uwzględnia dodatkowo elementy „fantomowe" reprezentujące elipsy i charakterystyczne dla języka rosyjskiego puste konstrukcje łącznikowe czasu teraźniejszego, np. *Ja tancior.* (*Jestem tancerzem*).

Zdania składające się na bank drzew są wpierw poddawane automatycznej analizie morfosyntaktycznej i syntaktycznej za pomocą narzędzi opracowanych dla systemu NLP ETAP-3 ukierunkowanemu głównie na tłumaczenie maszynowe automatycznego (Apresjan *et al.*, 1992, 2003), a następnie ręcznie ujednoznaczniane na obu tych poziomach.

[13]Można uznać, że poszczególne znaczniki wiążą leksemy nimi oznakowane. Ma to sens dla znaczników taksonomicznych i topologicznych. Jednak w wypadku znaczników mereologicznych i derywacyjnych informacja, że dwa leksemy są częściami czy też zdrobnieniami niewiele nam mówi, skoro nie wiemy, czego są częścią i zdrobnieniem.

SynTagRus kodowany jest w XML-u zgodnie z standardem TEI (por. sekcja 2.2). Podział na zdania oznaczany jest jako element `<S>`, segmenty oznaczane są jako `<W>`, każdy z nich posiada jednoznaczny identyfikator `ID`. Dwa atrybuty, `LEMMA` i `FEAT`, służą do przypisywania segmentom opisu morfosyntaktycznego. Kolejne dwa atrybuty segmentu przeznaczone są do przechowywania właściwej struktury syntaktycznej. Pierwszym z nich jest `DOM` zawierający identyfikator segmentu, od którego opisywany segment zależy, drugim jest `LINK` zawierający etykietę funkcji syntaktycznej.

4.4 Czeski Korpus Narodowy

Bardzo znanym i często cytowanym korpusem, traktowanym jako wzorcowy wśród korpusów języków słowiańskich, jest *Czeski Korpus Narodowy* (czes. *Český Národní Korpus*), `http://ucnk.ff.cuni.cz` (Čermák, 1997, 1998, 2001).

CNK składa się z szeregu zasobów związanych z etapami przetwarzania tekstu. Pierwszym zasobem jest Archiwum Korpusu zawierające teksty w różnych formatach, tak jak zostały one dostarczone. Kolejnym zasobem jest Bank Korpusu, zawierający teksty oczyszczone z fragmentów niejęzykowych (tabel itp.) i obcojęzycznych, a następnie przetworzone na jednolity format SGML. Teksty te opatrzone są bogatymi metadanymi.

Bank dzieli się na podkorpusy synchroniczny i diachroniczny (DIAKORP); pierwszy z nich zawiera z kolei korpus języka pisanego SYN2000, języka mówionego ORAL-PMK i ORAL-BMK oraz dialektów DIALKORP. Granicę pomiędzy korpusem synchronicznym i diachronicznym stanowi przełom 1989/1990, z wyjątkiem najważniejszych, wciąż wznawianych i czytanych utworów literatury pięknej.

Bank jako całość zawierał w 2000 roku ok. 330 mln. wyrazów, jego podstawowa, zrównoważona część — SYN2000 — 100 mln. wyrazów. Dwa korpusy języka mówionego, zebrane w Pradze i Brnie, zawierają łącznie ok. 800 tys. wyrazów. Korpus jest zrównoważony pod względem płci, wieku i wykształcenia mówiących. Korpus diachroniczny zawiera ok. 2 mln. wyrazów.

Struktura CNK (SYN2000) przedstawiona jest w Tabeli 10. Daje się zauważyć niski udział beletrystyki w stosunku do innych korpusów, np. BNC (27%) czy NKRJ (40%) przy wysokim udziale tekstów prasowych. Była to świadoma decyzja twórców korpusu, poprzedzona badaniami czytelnictwa.

Czeski Korpus Narodowy jest znakowany morfosyntaktycznie. Struktura znaczników ma charakter pozycyjny, tzn. pozycja w znaczniku determinuje interpretację danej cechy morfosyntaktycznej.

styl tekstu	udział	styl tekstu	udział
Teksty specjalistyczne	25%	sztuka	3%
ekonomia i zarządzanie	2%	technologia	5%
filozofia i religie	1%	życie codzienne	6%
nauki przyrodnicze	3%		
nauki społeczne	4%	**teksty prasowe**	60%
prawo i bezpieczeństwo	1%	**beletrystyka**	15%

Tabela 10: Struktura Czeskiego Korpusu Narodowego

5 Korpus IPI PAN

Poniższy rozdział poświęcony zostanie Korpusowi IPI PAN (dalej: Kipi) stworzonemu w Zespole Inżynierii Lingwistycznej IPI PAN w ramach grantu KBN 7 T11C043 20. (Przepiórkowski, 2004; Przepiórkowski *et al.*, 2003).

5.1 Reprezentacja Korpusu IPI PAN

Utwory zawarte w Korpusie IPI PAN dostarczone były w różnych stronach kodowych i formatach (por. sekcja 2.2). Wszystkie te utwory sprowadzane są do jednolitej postaci. Podobnie jak korpusy BNC (sekcja 3.2) i ANC (sekcja 3.3), teksty Kipi reprezentowane w jednolitym formacie, będącym nieznacznie zmienioną postacią XCES (por. sekcja 2.2), zaś przyjętą stroną kodową jest UTF-8. Kipi jest drzewem katalogów (niektóre utwory są grupowane, np. dzieła tego samego autora bądź artykuły tej samej gazety, dzielone następnie na poszczególne wydania). Poszczególne utwory znajdują się w liściach tego drzewa. Każdy utwór reprezentowany jest za pomocą trzech plików:

- `header.xml` zawierającego metadane,
- `text.xml` zawierającego tekst wraz z informacją strukturalną (podział na rozdziały, akapity itp.),
- `morph.xml` zawierającego tekst wraz ze znakowaniem morfosyntaktycznym.

Utwory konwertowane były do postaci plików `header.xml` i `text.xml` półautomatycznie, tzn. po automatycznej konwersji następowała jej ręczna weryfikacja. Uzyskane teksty nie były normalizowane pod kątem sposobu reprezentacji liczb itp.

Pliki `morph.xml` tworzone już były w pełni automatycznie. Operacja ta dokonywana była w dwóch etapach. W pierwszym za pomocą prostego algorytmu wydzielane były w tekście zdania (por. Przepiórkowski, 2004). W drugim kroku w procesie znakowania morfosyntaktycznego wyznaczane były segmenty, wraz z przypisaniem właściwej interpretacji morfosyntaktycznej. Mianem segmentów określamy w ogólności ciąg znaków znajdujący się pomiędzy dwoma separatorami. Jest nimi większość znaków interpunkcyjnych oraz oczywiście spacja. Jednak czasem znak interpunkcyjny znajduje się wewnątrz segmentu (np. przy odmianie wyrazów obcych (*Lagrange'a*) i skrótowców (*PKS-em*)). Istnieje jednak także sytuacja przeciwna. Wśród klas fleksyjnych występuje tzw. aglutynant. Jest to specjalna forma czasownika, odmienna przez liczbę i osobę, która determinuje osobę czasownika w czasie przeszłym. W przyjętym rozwiązaniu bowiem formy przeszłe czasownika (zwane pseudoimiesłowami) występują wyłącznie w osobie trzeciej, zaś aglutynanty stanowią odrębne segmenty do nich przyłączone. Wynika to z faktu, że aglutynanty mogą odłączać się od czasowników i przyłączać do innych części mowy (*wyście, tum, żeśmy, byś*). Dotyczy to także partykuł (zwanych wraz z wykrzyknikami kublikami) *że, by*, które są traktowane jako osobne segmenty niezależnie od tego, czy występują samodzielnie, czy razem z innymi wyrazami (*gdzieżby(m), zrobiłby(m), zróbże*).

Proces znakowania morfosyntaktycznego także przebiega dwufazowo. Najpierw analizator morfologiczny *Morfeusz* (Woliński, 2006) przypisuje każdemu segmentowi wszystkie możliwe interpretacje, zgodnie z tagsetem IPI PAN (Przepiórkowski i Woliński, 2003a,b). Następnie najbardziej prawdopodobna spośród tych interpretacji jest wybierana przez tager. Dwa tagery: prosty trigramowy tager stochastyczny (Dębowski, 2003, 2004) oraz regułowy (Piasecki i Gaweł, 2005; Broda *et al.*, 2008) zostały wytrenowane specjalnie do znakowania Kipi. Poza oznaczeniami występującymi w *Morfeuszu*, w Kipi znajduje się pewien specjalny znacznik ign, opisujący wyraz nieznany analizatorowi morfologicznemu[14].

Dokładniejsze omówienie Korpusu IPI PAN znajduje się w (Przepiórkowski, 2004).

5.2 Struktura korpusu

Korpus IPI PAN tworzony był metodą „oportunistyczną", co oznacza, że większość utworów, jakie udało się zgromadzić w postaci elektronicznej[15] wraz z wymaganymi prawami autorskimi (korpus jest udostępniany publicznie) została

[14]Wyrazy takie mogą być zinterpretowane w podkorpusie oznakowanym ręcznie. W szczególności, zostały w tym celu wprowadzone dwa dodatkowe znaczniki xxx i xxs oznaczające odpowiednio wyrażenia obcojęzyczne i takież frazy nominalne.

[15]Funduszy na drogi proces przekształcania tekstów z postaci drukowanej na elektroniczna (tzw. OCR-owanie) nie było.

w nim umieszczona. Utwory były wprowadzane do korpusu w całości, bez losowania mniejszych próbek. W ten sposób powstał korpus duży (ok. 250 mln. segmentów), lecz nie zrównoważony ani pod względem stylów tekstów, ani pod względem rozmiaru rozważanych w nim utworów. Dane statystyczne dotyczące Kipi przedstawione zostały w Tabeli 11. W pierwszej kolumnie tabeli widnieją style tekstów (wraz z ich podstylami). Podstyle posiadają style urzędowo-kancelaryjny i naukowo-dydaktyczny oraz potencjalnie artystyczny, choć w Kipi jedynym jego podstylem jest proza. W kolejnych trzech kolumnach występuje liczba plików danego stylu, liczba występujących w nim segmentów oraz średnia liczba segmentów na plik. Jak widać, średnio najdłuższe są utwory z literatury faktu (97593 segmenty), a najkrótsze — artykuły prasowe (403 segmenty). Jest oczywiste, że im większy udział w danym stylu mają publikacje książkowe, tym większa jest średnia długość utworu. W korpusie występuje tylko jeden utwór zaklasyfikowany jako potoczny.

Kolejna kolumna tabeli (oznaczona jako *%seg.*) pokazuje faktyczną strukturę korpusu, czyli udział poszczególnych stylów w całości korpusu, liczony jako procent segmentów. Dominacja dwóch stylów, publicystycznego i urzędowo-kancelaryjnego (a w tym przede wszystkim protokołów sejmowych) jest miażdżąca. Należy przy tym zauważyć, że styl publicystyczny jest sam w sobie dość zróżnicowany; znajdują się w nim informacje bieżące (nie tylko polityczne), lokalne, sportowe oraz utwory publicystyczne. Natomiast styl urzędowo-kancelaryjny jest mieszaniną tekstów sformalizowanych (ustawy, ale też protokolarne elementy posiedzeń sejmowych) i mówionych „oficjalnych" (przemówienia poselskie).

Ostatnie trzy kolumny tabeli pokazują liczbę specjalnych segmentów, tzn. aglutynantów i znaków interpunkcyjnych, oraz ich procentowy udział w utworach danego stylu. Jak widać, te segmenty nie będące samodzielnymi wyrazami stanowią około 1/5 wszystkich segmentów.

Choć rozmiar korpusu wydaje się wielki, to brak zrównoważenia zmniejsza wagę tego faktu. Korpus IPI PAN nie jest też korpusem synchronicznym, choć trudno byłoby uznać go za korpus diachroniczny, gdyż żaden porządek utworów względem czasu nie został w nim uwzględniony. W Tabeli 12 przedstawiona jest struktura korpusu pod kątem liczby utworów napisanych i opublikowanych w danym okresie czasu. Jak widać, artykuły prasowe nie mają wpisanej do pliku nagłówkowego daty publikacji, choć data taka powinna być dostępna w innej postaci. Wszystkie teksty prasowe oraz urzędowo-kancelaryjne pochodzą z okresu 1990–2006. Utwory napisane przed 1939 rokiem to klasyka literatury pięknej. Tak duży jej udział w korpusie spowodowany jest m.in. wygaśnięciem praw autorskich dla tych dzieł. Data publikacji znana jest dla większej liczby utworów niż data ich napisania.

W Kipi zostały wydzielone także dwa mniejsze podkorpusy:

Tabela 11: Struktura Korpusu IPI PAN

Styl tekstu	l. plik.	l. segment.	średnia	% seg.	l. aglut.	l. interp.	(%)
potoczny	1	31459	31459	0,02	399	7801	26,07
artystyczny (proza)	74	5927448	80100	2,33	47618	1241571	21,75
literatura faktu	7	683153	97593	0,27	3546	126724	19,07
publicystyczny	353817	142519621	403	55,99	401393	28319848	20,15
urzędowo-kancelaryjny	3431	102836825	29973	40,40	326766	15263712	15,16
protokół	1598	94589984	59192	37,16	326719	13772187	14,91
ustawa	1833	8246841	4499	3,24	47	1491525	18,09
informacyjno-poradnikowy	2	93818	46909	0,04	464	16810	18,41
naukowo-dydaktyczny	51	2432300	47692	0,96	3469	475884	19,71
podręcznik	2	90389	45194	0,04	140	15231	17,01
popularno-naukowy	6	208474	34746	0,08	178	35187	16,96
humanistyczny	26	1808652	69563	0,71	3117	363544	20,27
przyrodniczy	8	214629	26828	0,08	23	40327	18,80
techniczny	9	110156	12239	0,04	11	21595	19,61
RAZEM	357383	254524624	712	100,00	783655	45452350	18,17

data	napisania	publikacji	data	napis.	publik.
1877–1910	35	34	2001	89	239
1920–1939	8	11	2002	121	636
1950–1989	31	92	2003	127	318
1990–1995	427	539	2004	48	376
1996–2000	559	781	2005	1	105
niezdefin.	355936	354244	2006	1	8
RAZEM	357383	357383			

Tabela 12: Podział Korpusu IPI PAN na okresy czasu

- Podkorpus zrównoważony zawierający ok. 30 mln. segmentów, stanowiący podzbiór tekstów zrównoważony pod względem stylu; jego struktura widnieje w Tabeli 13;
- podkorpus oznakowany ręcznie, zawierający ok. 1 mln segmentów.

styl	udział	styl	udział
publicystyka	50%	proza współczesna	10%
protokoły parlamentarne	15%	proza starsza	10%
ustawy	5%	literatura faktu	10%

Tabela 13: Struktura podkorpusu zrównoważonego Kipi

Z tabeli 13 wynika, ze utwory ze stylu naukowo-dydaktycznego nie zostały w podkorpusie zrównoważonym uwzględnione.

6 Narodowy Korpus Języka Polskiego

W chwili zakończenia prac nad niniejszą książką, Kipi nie jest już największym korpusem języka polskiego. Jego miejsce zajął *Narodowy Korpus Języka Polskiego* (NKJP, `http://nkjp.pl`) stworzony w ramach grantu rozwojowego MNiSW R17 003 03 jako połączenie prac korpusowych prowadzonych w IPI PAN, IJP PAN, ZJKiK Uniwersytetu Łódzkiego i Wydawnictwie Naukowym PWN (Przepiórkowski *et al.*, 2008, 2009, 2010). Przewyższa on pod każdym względem (rozmiaru, poziomów znakowania, metod znakowania i standardów przyjętych podczas ich realizacji) swojego poprzednika. Doświadczenia zdobyte

podczas tworzenia Kipi oraz prac korpusowych pozostałych uczestników konsorcjum zostały wykorzystane podczas jego budowy, co ma korzystny wpływ na jego jakość.

NKJP w całości zawiera 1,8 mld. segmentów, z czego 300 mln. segmentów stanowi korpus zrównoważony, co stanowi znaczący wzrost w stosunku do Kipi. Szczególną wartość ma podkorpus swobodnego języka mówionego wielkości ok. 3 mln. segmentów, powstały na gruncie podkorpusu języka mówionego korpusu PELCRA stworzonego w ZJKiK UŁ.

Tagset zastosowany do znakowania morfosyntaktycznego niewiele różni się od wykorzystanego w Kipi(por. Przepiórkowski, 2009). Pozostałe poziomy znakowania są zupełną nowością na gruncie polskiej lingwistyki korpusowej.

Ręczne znakowanie podkorpusu ok. 1 mln segmentów, złożonego z losowo dobranych próbekpróbka z podkorpusu zrównoważonego, zrealizowane zostało za pomocą zmodyfikowanej wersji programu *Anotatornia* (Przepiórkowski i Murzynowski, 2009). Prace te obejmowały także pilotażowe znakownie wszystkich wystąpień 100 wybranych wieloznacznych wyrazów ich sensami.

Kolejny poziom znakowania obejmuje jednostki identyfikacyjne (ang. *named entities*), w tym wieloczłonowe (Savary *et al.*, 2010; Savary i Piskorski, 2009). Znakowanie obejmuje nazwy osobowe, geograficzne, organizacji i instytucji, wyrażenia temporalne oraz określenia pochodne (np. *warszawski, warszawiak*). Znakowanie jednostek wieloczłonowych nie jest ograniczone do jednostek o maksymalnej długości, lecz ma charakter hierarchiczny (np. *[ulica [Adama Mickiewicza]]*). Znakowanie obejmuje jednostki nieciągłe.

Natomiast znakowanie syntaktyczne (Głowińska i Przepiórkowski, 2010) ma charakter powierzchniowy i obejmuje dwa poziomy: wyrazów syntaktycznych i grup (fraz) syntaktycznych. Wyrazy syntaktyczne zazwyczaj pokrywają się z segmentami, obejmują jednak także analityczne formy czasu i trybu, czasowniki z się morfologicznym, spójniki i przyimki złożone itp. Wyrazy syntaktyczne mogą być nieciągłe (*będę szybko szedł*) oraz mogą na siebie nachodzić *będę szedł i śpiewał, bał się zaśmiać*. Do znakowania wyrazów syntaktycznych opracowany został odrębny tagset. Drugi poziom obejmuje grupy syntaktyczne, w tym nieciągłe.

7 Banki drzew

Jak wspominaliśmy na początku poprzedniego rozdziału, korpusy tekstów stanowią dość rozległą i różnorodną klasę zasobów, w zależności od przeznaczenia i sposobu znakowania. Ich specyficzną podklasę stanowią *banki drzew* (ang. *treebanks*), które poza standardowym znakowaniem morfosyntaktycznym zawierają informację składniową. Informacja ta ma zazwyczaj postać drzew rozbioru składniowego, stąd nazwa.

7.1 Bank drzew Penn

Jednym z najbardziej znanych i cytowanych banków drzew jest Penn (ang. *Penn Treebank*) stworzony na uniwersytecie Pensylwanii we współpracy z innymi instytucjami. Jego pierwsza wersja powstała w latach 1989–1992 zawierała ok. 4,5 mln. wyrazów angielszczyzny amerykańskiej (Marcus, 1994; Marcus *et al.*, 1993; Taylor *et al.*, 2003).

7.1.1 Struktura korpusu

Pierwsza wersja banku drzew Penn zawiera ok. 2,9 mln. wyrazów (segmentów).[16] Struktura korpusu przedstawiona jest w Tabeli 14 pochodzącej z (Marcus *et al.*, 1993). W szczególności, bank zawiera w całości korpus Browna, którego struktura przedstawiona została w Tabelach 1 i 2 na stronie 12. Brak danych dotyczących liczby próbek dla poszczególnych źródeł tekstu i wielkości tychże.

źródło tekstu	l. segmentów
Streszczenia Departamentu Energii	231 404
Doniesienia Dow Jones Newswire	1 061 166
Biuletyny Departamentu Rolnictwa	78 555
teksty Biblioteki Amerykańskiej	105 652
transkrypcje radia WBUR	11 589
wiadomości MUC-3	111 828
zdania z Manuala IBM	89 121
zdania z ATIS	19 832
korpus Browna	1 172 041
RAZEM	2 881 188

Tabela 14: Skład banku drzew Penn

Strukturę banku drzew Penn można by uznać za zrównoważoną, gdyby nie dominacja Dow Jones Newswire (korpus Browna jest sam w sobie zrównoważony). Także teksty niektórych pozostałych stylów (np. Manuala IBM) są zbyt jednolite ze względu na pojedyncze źródło pochodzenia, ale ze względu na swój rozmiar mają mniej dominujący charakter.

[16] Liczba segmentów oznakowanych wyłącznie morfosyntaktycznie wynosi ok. 4,9 mln., a różnica ta dotyczy wyłącznie doniesień Dow Jones Newswire.

7.1.2 Znakowanie morfosyntaktyczne

Znakowanie morfosyntaktyczne banku drzew Penn oparte zostało na zestawie znaczników wprowadzonym w korpusie Browna (sekcja 3.1), jednak został on znacznie uproszczony w celu zmniejszenia rozproszenia danych, co ma szczególne znaczenia podczas ich statystycznego przetwarzania. Na przykład zrezygnowano z odrębnych znaczników dla czasowników *be, do, have* wskazując przy tym na ich odtwarzalność na podstawie danych leksykalnych. Zrezygnowano także z rozróżnienia zaimków znajdujących się na pozycji podmiotu bądź dopełnienia (*he, him*) nawet wówczas, gdy nie istnieje różnica na poziomie leksykalnym (*you*), ze względu na odtwarzalność ze struktury syntaktycznej (Marcus *et al.*, 1993). Faktycznie, odtwarzalność leksykalna zdawałaby się tu mało pomocna, gdyż jest to ukrywanie pewnej informacji w słowniku. Rozumowanie takie nie bierze pod uwagę przydatności takiej informacji podczas automatycznego tworzenia struktury syntaktycznej, gdy z natury rzeczy nie jest ona znana. Mimo to podejście takie wydaje się uzasadnione, skoro takie rozróżnienie nie istnieje dla rzeczowników występujących na tych samych pozycjach.

Redukcja zestawu znaczników miała także na celu zwiększenie spójności znakowania. Na przykład, przysłówki *there, now* były jednoznacznie uznawane za „zwykłe" przysłówki (ang. *adverbs*), zaś ich odpowiedniki *here, then* mogły być dodatkowo oznaczane jako przysłówki rzeczownikowe (ang. *nominal adverbs*).

W banku Penn, podobnie jak w Korpusie Browna, znakowanie morfosyntaktyczne wyrazów zależy od ich funkcji syntaktycznej. W banku Penn dopuszcza się dodatkowo niejednoznaczne znakowanie wyrazów, gdy kontekst nie daje dostatecznych podstaw do ujednoznacznienia.

Bank drzew Penn został wstępnie automatycznie oznakowany morfosyntaktycznie za pomocą tagera stochastycznego PARTS (Church, 1988) działającego na zbliżonym zbiorze znaczników. Wyniki zostały automatycznie przekształcone na zbiór znaczników Penn, co nieco powiększyło poziom błędu. Następnie korpus został ponownie oznakowany kaskadą tagerów regułowych i stochastycznych opracowanych na podstawie wcześniejszych danych (już specjalnie dla znaczników Penn); ostateczny poziom błędu wyniósł 2–6%.

Uzyskany w ten sposób oznakowany korpus został następnie poddany ręcznej korekcie, przy czym Anotatorzy mogli zmieniać oznaczenia jedynie w ramach dopuszczalnych dla danego wyrazu. Podejście to zostało wpierw eksperymentalnie sprawdzone (Marcus *et al.*, 1993), poprzez poddanie pewnej próbki całkowicie ręcznemu znakowaniu. Metoda ta okazała się dwukrotnie gorsza pod względem szybkości, zgodności miedzy anotatorami oraz uzyskanej dokładności (ang. *accuracy*).

7.1.3 Znakowanie syntaktyczne

Marcus *et al.* (1993) nazywają znakowanie syntaktyczne *nawiasowaniem*, gdyż drzewa rozbioru reprezentowane są w nawiasowej postaci listowej. Każdy nawias otwierający etykietowany jest znacznikiem syntaktycznym określającym typ reprezentowanej frazy. Zestaw znaczników syntaktycznych prezentowany jest w Tabeli 15. Znaczniki `*, 0, T, NIL` oznaczają elementy puste, które w tekście tak naprawdę nie występują.

ozn.	opis	ozn.	opis
NP	fraza nominalna	S	proste zdanie oznajm.
VP	fraza czasownikowa	WHNP	fraza względna nomin.
ADJP	fraza przymiotn.	WHADVP	fraza względna przysł.
ADVP	fraza przysłówkowa	WHPP	fraza względna przyim.
PP	fraza przyimkowa	X	składnik niezn. kateg.
SBAR	zdanie wprow. przez spójnik podrz.	*	podmiot domyślny zdań bezokoliczn. i rozkazuj.
SBARQ	pyt. wprow. przez zaimek pytajny	0	pusty wariant *that* w zdaniach podrzędnych
SINV	zdanie oznajmujące z inwersją podmiotu	T	wskaźn. położenia zaimka w zdaniu względnym
SQ	podrz. SBARQ (bez zaimka pytajn.)	NIL	wskaźn. położenia fr. rzecz. za zaimkiem

Tabela 15: Znaczniki syntaktyczne w banku drzew Penn

Proces znakowania syntaktycznego przebiegał analogicznie do procesu znakowania morfosyntaktycznego, tzn. polegał na ręcznej korekcie wyników uzyskanych automatycznie. Wstępnej analizy syntaktycznej dokonano przy użyciu deterministycznego parsera Fidditch (Hindle, 1989). Fidditch zawsze dostarcza dokładnie jeden rozbiór dowolnego zdania. Jeśli funkcja danego wyrażenia w zdaniu nie daje się ustalić jednoznacznie, parser tworzy ciąg poddrzew, formując jedynie częściową strukturę zdania. Tak więc można uznać Fidditch za rodzaj parsera hybrydowego, który łączy w sobie cechy parsera głębokiego i powierzchniowego.[17] W szczególności, parser ten n i g d y nie dowiązuje fraz przyimkowych, zdań względnych i modyfikatorów przysłówkowych do żadnych struktur nadrzędnych.

[17]Parser głęboki tworzy pełen (głęboki) rozbiór zdania. Parser powierzchniowy jedynie wyznacza w zdaniu granice poszczególnych fraz, nie tworząc jego pełnej struktury.

```
(S
  (NP Battle-tested/NNP industrial/JJ managers/NNP here/RB)
  always/RB
  (VP buck/VB up/IN
      (NP nervous/JJ newcomers/NNS)
      (PP with/IN
          (NP the/DT tale/NN
              (PP of
                  (NP (NP the/DT
                          (ADJP first/JJ
                                (PP of/IN
                                    (NP their/PP$ countrymen/NNS)))
                          (S (NP *)
                             to/TO
                             (VP visit/VB
                                 (NP Mexico/NNP))))
                      ,
                      (NP (NP a/DT boatload/NN
                              (PP of/IN
                                  (NP (NP warriors/NNS)
                                      (VP-1 blown VBN
                                            ashore/RB
                                            (ADVP (NP 375/CD
                                                      years/NNS)
                                                  ago/RB)))))
                          (VP-1 *pseudo-dowiązanie*)))))))
.)
```

Rysunek 1: Przykładowy rozbiór zdania w banku drzew Penn

Dobre pokrycie gramatyczne (czysto syntaktyczne, bez wykorzystywania jakiejkolwiek informacji semantycznej bądź pragmatycznej) języka angielskiego przez Fidditch powodowało, że Anotatorzy nie musieli korygować zbyt wielu drzew, które zostały błędnie stworzone dla danego zdania. Ich zadaniem była raczej prawidłowe łączenie poddrzew tam, gdzie parser pozostawił analizę powierzchniową.

Zadanie korekty znakowania syntaktycznego jest znacznie trudniejsze od korekty znakowania morfosyntaktycznego i zajmuje znacznie więcej czasu tak na etapie uczenia się, jaki i realizacji. Dlatego twórcy banku Penn zdecydowali się na uproszczenie struktury drzew rozbioru w stosunku do tej tworzonej przez Fidditch, przede wszystkim poprzez usunięcie węzłów, w których drzewo nie

ulega rozgałęzieniu. Uproszczenie takie także dokonywane było automatycznie. Przykład (pochodzący z Marcus *et al.*, 1993) rozbioru zdania *Battle-tested industrial managers here always buck up nervous newcomers with the tale of the first of their countrymen to visit Mexico, a boatload of warriors blown ashore 375 years ago* (*Zaprawieni w boju tutejsi menadżerowie przemysłowi zawsze zabawiają przejętych nowicjuszy opowiastką o ich pierwszych krajanach wizytujących Meksyk, na statku pełnym żołnierzy przygnanym do brzegu 375 lat temu.*) zaprezentowany jest na Rysunku 1. Szczegółowe zasady znakowania syntaktycznego banku Penn można znaleźć w (Santorini i Marcinkiewicz, 1991).

W późniejszej fazie tworzenia korpusu (Taylor *et al.*, 2003) powyższe szkieletowe znakowanie składniowe zostało uzupełnione o informacje predykatywno-argumentowe, co prowadzi do odróżnienia argumentów predykatów od fraz luźnych. Każdy argument został opatrzony etykietą semantyczną typu podmiot, dopełnienie itp. Wprowadzono też mechanizmy kontekstowe służące do wyznaczania konstrukcji nieciągłych oraz powiązania elementów pustych, takich jak podmiot frazy bezokolicznikowej czy konstrukcje względne, z odpowiednimi elementami struktury zdania.

7.2 Niemieckojęzyczne banki drzew

Sporo banków drzew opracowano dla języka niemieckiego. Najbardziej znane z nich to *Negra* i *TiGer*.

Negra (Skut *et al.*, 1998; Brants *et al.*, 2003) to anotowany syntaktycznie korpus tekstów prasowych pochodzących z *Frankfurter Rundschau* opracowany na uniwersytecie Saarland. Wersja 2 korpusu zawiera 20 602 zdania (355 096 wyrazów). Jest on oznakowany morfosyntaktycznie (za pomocą tagsetu STTS — Stuttgart-Tübingen-Tagset) oraz syntaktycznie (Brants i Plaehn, 2000; Brants, 2000; Brants *et al.*, 1999; Brants i Skut, 1998). Pełne znakowanie morfosyntaktyczne posiada jedynie 60 tys. tokenów; pozostałe znakowane są jedynie klasami gramatycznymi.

Znakowanie syntaktyczne zawiera struktury wolne od kontekstu (ang. *context-free structures*) z krzyżującymi się gałęziami (Skut *et al.*, 1997). Gałęzie te mogą być przekształcone w ścieżki w taki sposób, by uzyskać format banku drzew Penn. Wewnętrzna reprezentacja korpusu ma postać bazy danych SQL, jest ona przekształcana na postać zewnętrzną(Brants, 1997a,b). Postać ta nie jest plikiem XML, lecz ma kształt zbioru tabel.

Jeszcze chyba bardziej znany jest korpus *TiGer* (Brants *et al.*, 2002; Brants i Hansen, 2002). Ciekawe, że jest to także korpus tekstów prasowych pochodzących z *Frankfurter Rundschau*. Zawiera on ok. 900 tys. wyrazów składających

się na 50 tys. zdań, czyli jest ponad dwukrotnie większy od *Negry*. Jest on w całości znakowany morfosyntaktycznie (w sposób analogiczny do *NeGry*; (Plaehn i Brants, 2000)) i syntaktycznie (Dipper, 2000; König i Lezius, 2003; Zinsmeister *et al.*, 2001a,b). Korpus posiada dwa formaty: jeden tekstowy, zgodny z reprezentacją korpusu *NeGra*, drugi XML-owy (König i Lezius, 2000; Mengel i Lezius, 2000).

Korpus *TiGer* wyposażony jest w bogate narzędzie wyszukiwawcze *TiGerSearch* (Lezius i König, 2000; König *et al.*, 2003).

Bank drzew został także przekształcony na postać zależnościową (Forst *et al.*, 2004; By, 2009).

7.3 Francuskojęzyczne banki drzew

Jednym z bardziej znanych banków drzew dla języka francuskiego jest TALANA (Abeillé *et al.*, 1998; Abeillé i Clément, 1999).

Jest to korpus prasowy składający się ze zróżnicowanych tematycznie wyselekcjonowanych fragmentów *Le Mond* z lat 1989–1993. Po połączeniu form złożonych oraz pominięciu znaków przestankowych korpus zawiera ok. 870 tys. segmentów lematyzowanych do 17 tys. różnych lematów i składających się na 32 tys. zdań.

Podobnie jak w wypadku banku Penn (por. sekcja 7.1), proces znakowania banku TALANA przeprowadzany był dwufazowo: najpierw realizowane było znakowanie morfosyntaktyczne, a dopiero potem syntaktyczne. Znakowanie morfosyntaktyczne zawierało lematyzację, kategorie gramatyczne i ich podkategorie (np. rzeczowniki pospolite i własne, spójniki współrzędne i podrzędne) oraz oznaczenia fleksyjne.

Złożony charakter zestawu znaczników spowodował, że konieczne było wydzielenie jego podzbiorów: wąskiego dla tagera (w celu minimalizacji liczby błędów) i szerszego dla anotatorów. Zastosowano tager (Abeillé *et al.*, 1998) zbliżony do tagera Brilla (1993; 1993), tyle że wykorzystujący zewnętrzny słownik oraz wspomagający się zestawem ręcznie utworzonych reguł.

Znakowanie syntaktyczne korpusu ma charakter powierzchniowy, co zapewnia kompatybilność z różnymi formalizmami gramatycznymi. Zaznaczane są granice fraz, ich rodzaje i ew. podrodzaje oraz funkcje powierzchniowe (podmiot, dopełnienie itp.). Poza standardowymi rodzajami fraz takimi jak fraza rzeczownikowa NP, przyimkowa PP, przymiotnikowa AP, przysłówkowa AdP i zdaniowa SENT wyróżnione są jądro czasownikowe NP oraz frazy (zdania) finitywne S podzielone na podrzędne sub i relacyjne rel, zaś kategoria VP obejmuje frazy bezokolicznikowe inf oraz okolicznikowe part. Poza tymi podstawowymi

rodzajami fraz wyróżnione są jednostki wieloczłonowe: nazwy własne name, liczby number, daty date i tytuły title.

Proces znakowania ma charakter półautomatyczny. Wpierw wyznaczane są jednostki wieloczłonowe, za pomocą automatu skończeniestanowego powstałego na bazie zestawu ręcznie utworzonych wyrażeń regularnych. Automat błędnie wyznacza około 3% jednostek, zaś około 20% pozostawia nierozpoznanych. Niedobory te są ręcznie korygowane przez anotatorów.

Właściwa analiza syntaktyczna dokonywana jest przez opracowany specjalnie do tego zadania parser powierzchniowy (Clément i Kinyon, 2000; Kinyon, 2000), a jego wyniki są znów poddawane ręcznej korekcji. Parser w ograniczonym zakresie uwzględnia zagnieżdżenia fraz. Ostatni krok stanowi wyznaczenie funkcji powierzchniowych oraz walencji czasowników głównych.

```
<S><PP> Au_cours_de:P
        <NP> la:D-fs  conference_de_presse:NC-fs
             <Srel><NP>:SUJ qui:PROR-3fs</NP>
                  <VN> a:VP-3s  clos:VK-ms</VN>
                  <NP> cette:D-fs  rencontre:NC-fs</NP>
             </Srel>
   </NP></PP> ,:PONCT
   <NP> le:D-ms premier_ministre:NC-ms <AP> est-allemand:A-ms</AP></NP>
   <VN> est:VP-3s revenu:VK-ms</VN>
   <PP> sur:P  <NP> les:D-mp  incidents:NC-mp
        <PP> de:P  <NP> lundi:NC-ms  soir:NC-ms</NP></PP>
        <Srel><PP> au cours de:P  <NP> lesquels:PROR-3mp</NP></PP>
             <NP> des:D-mp  manifestans:NC-mp</NP>
             <VN> ont:VP-3p  mis a sac:VK-ms</VN>
             <NP>le:D-ms  siege:NC-ms  <AP> central:A-ms</AP>
                 <PP> de:P  <NP> la:D-fs  Stasi:NP-fs</NP></PP></NP>
   </Srel></NP></PP>
</S>
```

Rysunek 2: Przykładowy rozbiór zdania w banku drzew TALANA

Przykład (pochodzący z Abeillé *et al.*, 2003) rozbioru zdania *Au cours de la confèrence de presse qui a clos cette rencontre , le premier ministre est-allemand est revenu sur les incidents de lundi soir au cours desquels des manifestants ont mis à sac le siège central de la Stasi.* (*Podczas konferencji prasowej kończącej to spotkanie, wschodnioniemiecki premier mówił o incydentach z poniedziałku rano, podczas których manifestanci splądrowali siedzibę główną Stasi.* przedstawiony jest na Rysunku 2.

7.4 Praski zależnościowy bank drzew

Znanym i często wspominanym w literaturze bankiem drzew dla języków słowiańskich jest *praski zależnościowy bank drzew* (ang. *Prague Dependency Treebank*, PDT) Böhmová *et al.* (2003); Hajič (1998, 2005); Hajičová (1998, 1999); Hajičová *et al.* (2001). Został on opracowany dla tekstów wybranych z Czeskiego Korpusu Narodowego (por. sekcja 4.4). Podstawą formalną banku jest Generatywny Opis Funkcjonalny (ang. *Functional Generative Description*, FGD) (Sgall *et al.*, 1986; Hajičová *et al.*, 1998). Ostatecznym formatem zapisu PDT jest XML (Pajas i Štěpánek, 2005).

Zasadniczą cechą PDT jest wielopoziomowość znakowania. Pierwsze dwa poziomy (warstwy), morfosyntaktyczny (m-layer) i syntaktyczny (analityczny — a-layer) tworzą PDT 1.0.[18] Znakowanie syntaktyczne ma charakter zależnościowy, zgodnie z FGD. Każdy segment ma jednoznacznie przyporządkowaną parę ⟨lemat, znacznik⟩ na poziomie morfosyntaktycznym oraz parę ⟨nadrzędnik, funkcja analityczna⟩ na poziomie syntaktycznym. Przykładowe funkcje analityczne to Pred (predykat), Sb (podmiot), Obj (dopełnienie), Atr (atrybut), Adv (przysłówek) itd. W ten sposób powstaje struktura drzewa zależnościowego, którego węzłami są segmenty warstwy m, zaś krawędzie etykietowane są funkcjami analitycznymi. Korzeń drzewa stanowi dodatkowy, sztuczny węzeł.

Na podstawie ręcznego znakowania morfosyntaktycznego 1,8 mln. segmentów, wytrenowane i porównane zostały dwa tagery (Hajičová i Hladká, 1998; Hajič *et al.*, 2001). Co ciekawe, znakowanie ręczne było *a posteriori* poddawane automatycznej analizie morfologicznej, a niezgodność była podstawą do korekty zarówno znakowania, jak i analizatora.

Znakowanie analityczne 1,5 mln. segmentów dokonywane było ręcznie, przy czym adekwatna funkcja analityczna proponowana była przez niewielki zestaw ręcznie opracowanych reguł. W dalszej fazie prace wspomagane były przez parser statystyczny wytrenowany na danych uzyskanych wcześniej (Collins *et al.*, 1999).

Poziom morfosyntaktyczny uzupełniony został o znakowanie leksykalno-semantyczne segmentów należących do klas otwartych (autosemantycznych) sensami pochodzącymi z czeskiego wordnetu (Smrž, 2004). Do roku 2006 oznakowano ok. 20% PDT (Bejček *et al.*, 2006), przy czym znakowane były jedynie segmenty, których lematy zostały w czeskim wordnecie uwzględnione.

Dodanie warstwy tektogramatycznej (t-layer) doprowadziło do powstania PDT 2.0. Także na tym poziomie zdanie reprezentowane jest jako drzewo, jednak jego węzłami są wyłącznie wyrazy posiadające własną semantykę, wyrazy

[18] Pomijam tu warstwę zawierającą jedynie segmentację tekstu, w-layer.

funkcyjne (np. przyimki czy spójniki) są ignorowane. Węzłom warstwy tektogramatycznej przypisywane są odrębne „lematy tektogramatyczne". Zazwyczaj są to właściwe jednostki leksykalne, jednak na tym poziomie dodawane są także węzły „fikcyjne" nie mające odpowiedników na powierzchni, np. #PersPron dla zaimków osobowych, #EmpNoun dla elipsy rzeczownika, #Cor dla pustego podmiotu frazy bezokolicznikowej. Podobnie jak na poziomie analitycznym, węzłom warstwy `t` przyporządkowywana jest para ⟨nadrzędnik, funktor⟩. Funktory (podstawowe: ACT (agens), ADDR (odbiorca), PAT (obiekt), EFF (skutek) i ORIG (źródło) oraz pomocnicze: AIM (cel), BEN (beneficjent), DIR2 (kierunek przez), MANN (sposób) itp.) określają rolę semantyczną węzła względem nadrzędnika, czyli tak naprawdę etykietują krawędzie drzewa. Funktory nie zależą od wiązanych leksemów, jednak wyróżnienie ponad 30 oznacza przyjęcie w PDT wysokiego poziomu szczegółowości.

Większość opisu morfosyntaktycznego jest nieistotna z semantycznego punktu widzenia i jest ignorowana na poziomie tektogramatycznym. Nie mniej jednak część informacji, np. czas czasownika, stopień przymiotnika czy liczba rzeczownika ma wpływ na znaczenie wypowiedzenia (por. zdania *Piotr spotkał swojego najmłodszego brata.*, *Piotr spotyka swojego młodego brata.* i *Piotr spotka swoich młodszych braci.*), i powinna być przechowywana w węzłach tego poziomu (Razímová i Žabokrtský, 2005).

Do części PDT 2.0 dodane zostały relacje koreferencji segmentów (Kučová i Žabokrtský, 2005). Mogą nimi być także węzły „fikcyjne". W szczególności, jako rodzaj takiej relacji zostało uznane powiązanie pustego podmiotu frazy bezokolicznikowej (#Cor) z właściwym argumentem czasownika kontroli będącego nadrzędnikiem bezokolicznika.

8 Podsumowanie

Niniejszy raport zawiera przegląd najbardziej znanych korpusów tekstów. Wyszczególnione zostały najważniejsze ich cechy, tzn. rozmiar korpusu oraz poszczególnych próbek, struktura tematyczna, zakres i sposób znakowania. Opisane korpusy bardzo różnią się pod tym względem między sobą. Mniejsze korpusy, zazwyczaj starsze, np. korpus Browna, mają też mocno ograniczony rozmiar próbki. Trudniej jest porównywać strukturę i poziom zrównoważenia korpusów, gdyż sposób podziału korpusu na style jest charakterystyczną cechą korpusu, i kategorie te obejmują nieco odmienne obszary języka. Jedyną ewidentną charakterystyką jest obecność w korpusie tekstów mówionych oraz pochodzących z internetu. Specyficzną klasę stanowią korpusy prasowe. Ich popularność wynika z łatwości ich gromadzenia bez poważniejszej utraty pokrycia językowego[19].

[19] Teksty prasowe obejmują w praktyce wszystkie dziedziny życia.

Reguły segmentacji korpusów zostały ustandaryzowane i nie podlegają większym dyskusjom. Dotyczy to w szczególności korpusów kodowanych w XML-u zgodnie ze standardami TEI.

Sposób znakowania morfosyntaktycznego jest głęboko zależny od specyfiki języka, którego tekstu są znakowane, a także od tradycji językoznawczych w danym kraju, a nawet środowisku, który dany zestaw znaczników (tagset) opracowywało. Można tu wyróżnić dwie zasadnicze tendencje. Pierwsza polega na wprowadzeniu bogatego zestawu klas gramatycznych zróżnicowanych od razu pod kątem pewnych ich cech morfosyntaktycznych. Sztandarowym przykładem takiego podejścia jest tagset korpusu Browna. Druga polega na oddzieleniu klas gramatycznych od kategorii morfosyntaktycznych i wprowadzeniu znakowania dwupoziomowego. Każda klasa gramatyczna ma wówczas przypisany zestaw charakterystycznych kategorii. Taki sposób znakowania jest szczególnie charakterystyczny dla języków fleksyjnych.

Odrębny temat stanowi znakowanie syntaktyczne. W tym wypadku zależność od języka, przyjętego formalizmu gramatycznego, a nawet sposobu kodowania, jest jeszcze wyraźniejsza i trudna do uniknięcia. Dlatego istnieją banki drzew, w których ten sam zbiór zdań jest znakowany na kilka różnych sposobów (np. składnikowy i zależnościowy).

Niektóre korpusy posiadają także znakowanie semantyczne. Znakowanie takie może dotyczyć poszczególnych wyrazów, np. przy wykorzystaniu wordnetu (korpus SemCor) czy innych taksonomii (BNC, NKRJ, NKJP, PDT). Natomiast banki drzew mogą zawierać semantyczne znakowanie pewnych zależności między wyrazami, np. poprzez oznaczenie ról semantycznych (PDT) czy struktury funkcyjnej (Penn).

Bibliografia

A. Abeillé (red.) (2003) *Treebanks: Building and Using Parsed Corpora*, Kluwer Academic Publishers, Dordrecht, Holandia.

A. Abeillé, L. Clément (1999) *A tagged reference corpus for French*, w: *Proceedings of the LINC'99 Workshop at EACL'99*, Bergen, Norwegia.

A. Abeillé, L. Clément, R. Reyes (1998) *TALANA annotated corpus for French: the first results*, w: LREC (1998), s. 992–999.

A. Abeillé, L. Clement, F. ois Toussenel (2003) *Building a Treebank for French*, w: Abeillé (2003).

ACL (1998) *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational LinguisticsCOLING-ACL'98*, Montreal, Kanada.

J. Apresjan, I. Boguslavsky, L. Iomdin, B. Iomdin, A. Sannikov, V. Sizov (2006) *A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects*, w: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, s. 1378–1381, Genua, Włochy.

J. Apresjan, I. Boguslavsky, L. Iomdin, A. Lazursky, V. Sannikov, V. Sizov, L. Tsinman (2003) *ETAP-3 linguistic processor: A full-fledged NLP implementation of the MTT*, w: *Proceedings of the 1st International Conference on Meaning-Text Theory*, s. 279–288.

J. Apresjan, I. Boguslavsky, L. Iomdin, A. Lazursky, V. Sannikov, L. Tsinman (1992) *The Linguistics of a Machine Translation System*, Meta, t. 37, nr 1, s. 97–112.

G. Aston, L. Burnard (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edynburg, Wielka Brytania.

S. Atkins (1991) *Tools for computer-aided corpus lexicography: The Hector project*, Acta Linguistica Hungarica, t. 41, s. 5–72.

C. F. Baker, C. J. Fillmore, B. Cronin (2003) *The structure of the FrameNet database*, International Journal of Lexicography, t. 16, nr 3, s. 281–296.

C. F. Baker, C. J. Fillmore, J. B. Lowe (1998) *The Berkeley FrameNet Project*, w: ACL (1998), s. 86–90.

M. Bańko (red.) (2000) *Inny słownik języka polskiego*, Wydawnictwo Naukowe PWN, Warszawa.

E. Bejček, P. Möllerová, P. Straňák (2006) *Lexico-Semantic Annotation of PDT: Some Results, Problems and Solutions*, w: P. Sojka, I. Kopeček, K. Pala (red.), *Proceedings of the 9th International Conference on Text, Speech and Dialogue*, t. 4188 serii *Lecture Notes in Artificial Intelligence*, s. 21–28, Springer-Verlag, Brno, Czechy.

D. Biber (1988) *Variation across speech and writing*, Cambridge University Press, Nowy Jork, NY.

— (1995) *Dimension of register variation*, Cambridge University Press, Nowy Jork, NY.

— (2007) *Representativeness in corpus design*, w: Teubert i Krishnamurthy (2007), s. 134–165.

H. Bickel, M. Gasser, A. H. Buhofer, L. Hofer, C. Schön (2009) *Schweizer Text Korpus — Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten*, Linguistik online, t. 39, nr 3.

I. Boguslavsky, I. Chardin, S. Grigorjeva, N. Grigoriev, L. Iomdin, L. Kreidlin, N. Frid (2002) *Development of a dependency treebank for Russian and its possible applications in NLP*, w: LREC (2002), s. 852–856.

I. Boguslavsky, S. Grigorjeva, N. Grigorjev, L. Kreidlin, N. Frid (2000) *Dependency treebank for Russian: Concept, tools, types of information*, w: COLING (2000), s. 987–991.

A. Böhmová, E. Hajičová, J. Hajič, B. Hladká (2003) *The Prague Dependency Treebank: A three-level annotation scenario*, w: Abeillé (2003).

S. Brants, S. Dipper, S. Hansen, W. Lezius, G. Smith (2002) *The TIGER Treebank*, w: *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bułgaria.

S. Brants, S. Hansen (2002) *Developments in the TIGER Annotation Scheme and their Realization in the Corpus*, w: LREC (2002), s. 1643–1649.

T. Brants (1997a) *Internal and External Tagsets in Part-of-Speech Tagging*, w: *Proceedings of Eurospeech*, s. 2787–2790, Rodos, Grecja.

— (1997b) *The NEGRA Export Format*, CLAUS Report 98, Universität des Saarlandes, Computerlinguistik, Saarbrücken, Niemcy.

— (2000) *Inter-Annotator Agreement for a German Newspaper Corpus*, w: LREC (2000), s. 1435–1439.

T. Brants, O. Plaehn (2000) *Interactive Corpus Annotation*, w: LREC (2000), s. 453–459.

T. Brants, W. Skut (1998) *Automation of Treebank Annotation*, w: *Proceedings of New Methods in Language Processing NeMLaP–98*, s. 49–57, Sydney, Australia.

T. Brants, W. Skut, H. Uszkoreit (1999) *Syntactic Annotation of a German Newspaper Corpus*, w: *Proceedings of the ATALA Treebank Workshop*, s. 69–76, Paryż, Francja.

— (2003) *Syntactic Annotation of a German Newspaper Corpus*, w: Abeillé (2003).

E. Brill (1993) *A Corpus-Based Approach to Language Learning*, Rozprawa doktorska, University of Pennsylvania.

B. Broda, M. Piasecki, A. Radziszewski (2008) *Towards a set of general purpose morphosyntactic tools for Polish*, w: M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń (red.), *Proceedings of the Intelligent Information Systems XVI (IIS'08)*, Challenging Problems in Science: Computer Science, s. 441–450, Akademicka Oficyna Wydawnicza Exit, Zakopane.

L. Burnard (2007) *Where did we go wrong? A retrospective look at British National Corpus*, w: Teubert i Krishnamurthy (2007), s. 35–54.

T. By (2009) *The TiGer Dependency bank in Prolog Format*, w: M. A. Kłopotek, A. Przepiórkowski, S. T. Wierzchoń, K. Trojanowski (red.), *Recent Advances in Intelligent Information Systems*, Challenging Problems in Science: Computer Science, s. 119–129, Akademicka Oficyna Wydawnicza Exit, Warszawa.

J. B. Carroll, P. Davies, B. Richman (1971) *The American Heritage Word Frequency Book*, American Heritage Publishing Co., Nowy Jork, NY.

D. Ćavar, A. Geyken, G. Neumann (2000) *Digital Dictionary of the 20th Century German Language*, w: T. Erjavec, J. Gros (red.), *Proceedings of the Language Technologies Conference*, Ljubljana, Słowenia.

F. Čermák (1997) *Czech National Corpus: A case Study in Many Contexts*, International Journal of Corpus Linguistics, t. 2, s. 181–197.

— (1998) *Czech National Corpus: Its Character, Goal and Background*, w: Sojka *et al.* (1998), s. 9–14.

— (2001) *Language Corpora: The Czech Case*, w: Matoušek *et al.* (2001), s. 21–30.

K. Church (1988) *A stochastic parts program and noun phrase parser for unrestricted text*, w: *Proceedings of the 2nd ACL Conference on Applied Natural Language Processing (ANLP-88)*, s. 136–143, Austin, TX.

L. Clément, A. Kinyon (2000) *Chunking, marking and searching a morphosyntactically annotated corpus for French*, w: *Proceedings ACIDCA'2000*, Monastir, Tunezja.

COLING (2000) *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, Saarbrücken, Niemcy.

M. J. Collins, J. Hajič, E. Brill, L. Ramshaw, C. Tillmann (1999) *A Statistical Parser of Czech*, w: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, s. 397–404, College Park, MA.

S. Dipper (2000) *Grammar-based Corpus Annotation*, w: A. Abeillé, T. Brants, H. Uszkoreit (red.), *Proceedings of the Second Workshop on Linguistically Interpreted Corpora (LINC)*, s. 56–64, Luksemburg.

Ł. Dębowski (2003) *A reconfigurable stochastic tagger for languages with complex tag structure*, w: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2003)*, s. 63–70, Budapeszt, Węgry.

— (2004) *Trigram morphosyntactic tagger for Polish*, w: M. A. Kłopotek, S. T. Wierzchoń, K. Trojanowski (red.), *Proceedings of the Intelligent Information Systems New Trends in Intelligent Information Processing and Web Mining IIS:IIPWM'04*, Advances in Soft Computing, s. 409–413, Springer-Verlag, Zakopane.

T. Erjavec (2001) *The MULTEXT-East Resources Revisited*, ElsNews, t. 10, nr 3–2.

T. Erjavec, C. Krstev, V. Petkevič, K. Simov, M. Tadić, D. Vitas (2003) *The MULTEXT-East Morphosyntactic Specifications for Slavic Languages*, w: *Proceedings of the EACL 2003 Workshop on Morphological Processing of Slavic Languages*, s. 25–32, Budapeszt, Węgry.

C. Fellbaum (red.) (1998) *WordNet — An Electronic Lexical Database*, MIT Press, Cambridge, MA.

M. Filipenko, E. Paducheva, E. Rakhilina (1992) *Semantic dictionary viewed as a lexical database*, w: *Proceedings of the 14th International Conference on Computational Linguistics (COLING-1992)*, s. 1295–1299, Nantes, Francja.

C. J. Fillmore, N. Ide, D. Jurafsky, C. Macleod (1998) *An American National Corpus: A Proposal*, w: LREC (1998), s. 965–969.

C. J. Fillmore, C. R. Johnson, M. R. L. Petruck (2003) *Background to FrameNet*, International Journal of Lexicography, t. 16, nr 3, s. 235–250.

C. J. Fillmore, C. Wooters, C. F. Baker (2001) *Building a large lexical databank which provides deep semantics*, w: B. K. Tsou, O. Y. Kwong (red.), *Proceedings of the Pacific Asian Conference on Language, Information and Computation*, s. 3–25, Hong Kong.

M. Forst, N. Bertomeu, B. Crysmann, F. Fouvry, S. Hansen-Schirra, V. Cordoni (2004) *Towards a dependency-based gold standards of German parsers – The TiGer Dependency Bank*, w: *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora*, s. 31–37, Genewa, Szwajcaria.

W. N. Francis (2007) *Problems of assembling and computerizing large corpora*, w: Teubert i Krishnamurthy (2007), s. 285–298.

W. N. Francis, H. Kucera (1964, wersja poprawiona 1979) *Brown Corpus Manual*, Internet.

R. Garside (1996) *The Robust Tagging of Unrestricted Text: the BNC experience*, w: J. Thomas, M. Short (red.), *Using Corpora for Language for Language Reaserch: Studies in Honour of Geofrey Leech*, s. 167–180, Longman, Harlow.

E. Grishina, E. Rakhilina (2005) *Russian National Corpus (RNC): an overview and perspectives*, w: *Proceedings of the AATSEEL 2005.*

K. Głowińska, A. Przepiórkowski (2010) *The Design of Syntactic Annotation Levels in the National Corpus of Polish*, w: LREC (2010).

J. Hajič (1998) *Building a Syntactically Annotated Corpus*, w: E. Hajičová (red.), *Issues of Valency and Meaning*, s. 106–132, Charles University, Praga, Czechy.

— (2005) *Complex Corpus Annotation: The Prague Dependency Treebank*, w: M. Šimková (red.), *Insight into Slovak and Czech Corpus Linguistics*, s. 54–73, Veda, Bratysława, Słowacja.

J. Hajič, P. Krbec, P. Květoň, K. Oliva, V. Petkevič (2001) *Serial Combinations of Rules and Statistics: A Case Study in Czech Tagging*, w: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, s. 260–267, Tuluza, Francja.

E. Hajičová (1998) *The Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation*, w: Sojka *et al.* (1998), s. 45–50.

— (1999) *The Prague Dependency Treebank: Crossing the Sentence Boundary*, w: V. Matoušek, P. Mautner, J. Oceĺíková, P. Sojka (red.), *Proceedings of the 2nd International Workshop on Text, Speech and Dialogue*, s. 20–27, Springer-Verlag, Berlin.

E. Hajičová, J. Hajič, B. Hladká, P. Pajas, V. Řezničková, P. Sgall (2001) *The Current Status of the Prague Dependency Treebank*, w: Matoušek *et al.* (2001), s. 11–20.

E. Hajičová, B. Hladká (1998) *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset*, w: ACL (1998), s. 483–490.

E. Hajičová, B. H. Partee, P. Sgall (1998) *Topic-Focus Articulation, Tripartite Structures and Semantic Content*, t. 71 serii *Studies in Linguistics and Philosophy*, Kluwer Academic Publishers, Dordrecht, Holandia.

D. Hindle (1989) *Acquiring disambiguation rules from text*, w: *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89)*, s. 118–125, Vancouver, Kanada.

N. Ide (1998a) *Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora*, w: LREC (1998), s. 463–470.

— (1998b) *Encoding Linguistic Corpora*, w: *Proceedings of the Sixth Workshop on Very Large Corpora*, s. 9–17.

N. Ide, P. Bonhomme, L. Romary (2000) *XCES: AN XML-based Encoding Standard for Linguistic Corpora*, w: LREC (2000), s. 825–830.

N. Ide, R. Reppen, K. Suderman (2002) *The American National Corpus: More Than the Web Can Provide*, w: LREC (2002), s. 839–844.

N. Ide, K. Suderman (2004) *The American National Corpus First Release*, w: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, s. 1681–1684, Lisbon, Portugal.

S. Johansson, G. N. Leech, H. Goodluck (1978) *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*, Department of English, University of Oslo, Oslo, Norwegia.

A. Kinyon (2000) *Shallow parsing French using function words as triggers*, Rap. tech.

E. König, W. Lezius (2000) *A description language for syntactically annotated corpora*, w: COLING (2000), s. 1056–1060.

— (2003) *The TIGER language — A Description Language for Syntax Graphs, Formal Definition*, Rap. tech., Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

E. König, W. Lezius, H. Voormann (2003) *TIGERSearch User's Manual*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart, Niemcy.

W. Kopaliński (1968) *Słownik wyrazów obcych i zwrotów obcojęzycznych*, Wiedza Powszechna, Warszawa.

H. Kucera, W. N. Francis (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, RI.

I. Kurcz, A. Lewicki, J. Sambor, K. Szafran, J. Woronczak (red.) (1990) *Słownik frekwencyjny języka polskiego*, Instytut Języka Polskiego PAN, Kraków.

G. Kustova, O. Lashevskaja, E. Rakhilina, E. Paducheva (2007) *On Taxonomy in Cognitive Semantics and Corpus Linguistics: Parts of Body*, w: *Proceedings of the 10th International Cognitive Conference*, Kraków.

L. Kučová, Z. Žabokrtský (2005) *Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution*, w: Matoušek *et al.* (2005), s. 93–98.

O. Lashevskaja (2006) *Corpus-aided Construction Grammar: Semantic Tools in the Russian National Corpus*, w: *Proceedings of the 2th International Meeting of the German Cognitive Linguistic Association*, Monachium, Niemcy.

W. Lezius, E. König (2000) *Towards a search engine for syntactically annotated corpora*, w: W. Zühlke, E. G. Schukat-Talamazzini (red.), *Konvens 2000 Sprachkommunikation*, s. 113–116, VDE-Verlag, Ilmenau, Niemcy.

M. Liberman (1989) *Text on Tap: the ACL Data Collection Initiative*, w: *Proceedings of the DARPA Workshop on Speech and Natural Language*, s. 173–188, Morgan Kaufmann.

LREC (1998) *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC-1998)*, Grenada, Hiszpania.

— (2000) *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece.

— (2002) *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas, Hiszpania.

— (2010) *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, ELRA, Valetta, Malta.

H.-D. Maas (1996) *MPRO: Ein System zur Analyse und Synthese deutscher Wörter*, w: R. Hausser (red.), *Linguistische Verifikation, Sprache und Information*, 34, Max Niemeyer Verlag, Tybinga, Niemcy.

— (1998) *Multilinguale Textverarbeitung mit MPRO*, w: *Proceedings of the Europäische Kommunikationskybernetik heute und morgen*, Paderborn, Niemcy.

H.-D. Maas, C. Rösener, A. Theofilidis (2009) *Morphosyntactic and Semantic Analysis of Text: The MPRO Tagging Procedure*, w: C. Mahlow, M. Piotrowski (red.), *State of the art in computational morphology. Proceedings of the Workshop on systems and frameworks for computational morphology (SFCM 2009)*, t. 41 serii *Communications in computer and information science*, s. 76—87, Springer-Verlag.

M. Marciniak (red.) (2010) *Anotowany korpus dialogów telefonicznych*, Problemy Współczesnej Nauki. Teoria i Zastosowania: Inżyniera Lingwistyczna, Akademicka Oficyna Wydawnicza Exit, Warszawa.

M. P. Marcus (1994) *The Penn TreeBank: A revised corpus design for extracting predicate-argument structure*, w: *Proceedings of the ARPA Human Language Technology Workshop*, Morgan Kaufmann, Princeton, NJ.

M. P. Marcus, B. Santorini, M. A. Marcinkiewicz (1993) *Building a Large Annotated Corpus of English: The Penn Treebank*, Computational Linguistics, t. 19, nr 2, s. 313–330.

V. Matoušek, P. Mautner, R. Mouček, K. Taušer (red.) (2001) *Proceedings of the 4th International Conference on Text, Speech and Dialogue*, t. 2166 serii *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Zelezna Ruda, Czechy.

V. Matoušek, P. Mautner, T. Pavelka (red.) (2005) *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, t. 3658 serii *Lecture Notes in Artificial Intelligence*, Springer-Verlag, Karlovy Vary, Czechy.

I. Mel'čuk (1988) *Dependency Syntax: Theory and Practice*, State University of New York Press, Albany, NY.

I. Mel'čuk, A. Zholkovsky (1984) *Explanatory Combinatorial Dictionary of Modern Russian*, Wienner Slawistischer Almanach, Wiedeń, Austria.

A. Mengel, W. Lezius (2000) *An XML-based encoding format for syntactically annotated corpora*, w: LREC (2000), s. 121–126.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller (1990) *Introduction to WordNet: an on-line lexical database*, International Journal of Lexicography, t. 3, nr 4, s. 235–244.

G. A. Miller, C. Leacock, R. Tengi, R. Bunker (1993) *A semantic concordance*, w: *Proceedings of the ARPA Human Language Technology Workshop*, s. 303–308, Plainsboro, NJ.

S. M. Newman, R. W. Swanson, K. Knowlton (1959) *A Notation System for Transliterating Technical and Scientific Texts for Use in Data Processing Systems*, Rap. tech. 15, U.S. Department of Commerce.

J. Nivre, I. Boguslavsky, L. Iomdin (2008) *Parsing the SynTagRus Treebank*, w: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, s. 641–648, Manchester, Wielka Brytania.

P. Pajas, J. Štěpănek (2005) *A Generic XML-based Format for Structured Linguistic Annotation and its Application to Prague Dependency Treebank 2.0*, Rap. tech. TR-2005-29, ÚFAL MFF UK, Praga, Czechy.

M. Piasecki, B. Gaweł (2005) *A Rule-based Tagger for Polish Based on Genetic Algorithm*, w: M. A. Kłopotek, S. T. Wierzchoń, K. Trojanowski (red.), *Proceedings of the Intelligent Information Systems New Trends in Intelligent Information Processing and Web Mining IIS:IIPWM'05*, Advances in Soft Computing, s. 247–256, Springer-Verlag, Gdańsk.

O. Plaehn, T. Brants (2000) *Annotate – An Efficient Interactive Annotation Tool*, w: *Proceedings of the 6th ACL Conference on Applied Natural Language Processing (ANLP-2000)*, Seattle, WA.

A. Przepiórkowski (2004) *Korpus IPI PAN. Wersja wstępna*, Instytut Podstaw Informatyki, Polska Akademia Nauk, Warszawa.

— (2009) *A comparison of two morphosyntactic tagsets of Polish*, w: V. Koseska-Toszewa, L. Dimitrova, R. Roszko (red.), *Proceedings of the 4th MONDILEX Open Workshop on Representing Semantics in Digital Lexicography*, s. 138–144.

A. Przepiórkowski, P. Bański, Ł. Dębowski, E. Hajnicz, M. Woliński (2003) *Konstrukcja korpusu IPI PAN*, Polonica, t. XXII–XXII, s. 33–38.

A. Przepiórkowski, R. L. Górski, B. Lewandowska-Tomaszczyk, M. Łaziński (2008) *Towards the National Corpus of Polish*, w: *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, ELRA, Marrakech, Morocco.

A. Przepiórkowski, R. L. Górski, M. Łaziński, P. Pęzik (2009) *Recent developments in the National Corpus of Polish*, w: J. Levická, R. Grabík (red.), *Proceedings of the 5th International Conference on NLP, Corpus Linguistics, Corpus Based Grammar Research (Slovko 2009)*, s. 302–309.

— (2010) *Recent Developments in the National Corpus of Polish*, w: LREC (2010).

A. Przepiórkowski, G. Murzynowski (2009) *Manual Annotation of the National Corpus of Polish with Anotatornia*, w: S. Goźdź-Roszkowski (red.), *Practical Applications in Language Corpora (PALC'09)*, Peter Lang, Frankfurt nad Menem.

A. Przepiórkowski, M. Woliński (2003a) *A Flexemic Tagset for Polish*, w: *Proceedings of the Workshop of Morphological Processing of Slavic Languages, EACL-2003*, s. 33–40.

— (2003b) *A Morphosyntactic Tagset for Polish*, w: P. Kosta, J. Błaszczak, J. Frasek, L. Geist, M. Żygis (red.), *Investigations into Formal Slavic Linguistics*, s. 349–362, Peter Lang.

M. Razímová, Z. Žabokrtský (2005) *Morphological Meanings in the Prague Dependency Treebank 2.0*, w: Matoušek *et al.* (2005), s. 148–155.

R. Reppen, N. Ide (2004) *The American National Corpus: Overall goals and the first release*, Journal of English Linguistics, t. 32, nr 2, s. 105–113.

B. Santorini, M. A. Marcinkiewicz (1991) *Bracketing Guidelines for the Penn Treebank Project*, Rap. tech., Department of Computer and Information Science, University of Pennsylvania.

A. Savary, J. Piskorski (2009) *Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish*, w: M. A. Kłopotek, M. Marciniak, A. Mykowiecka, W. Penczek, S. T. Wierzchoń (red.), *Intelligent Information Systems*, Challenging Problems in Science: Computer Science, s. 141–154, Akademicka Oficyna Wydawnicza Exit, Warszawa.

A. Savary, J. Waszczuk, A. Przepiórkowski (2010) *Towards the Annotation of Named Entities in the National Corpus of Polish*, w: LREC (2010).

P. Sgall, E. Hajičová, J. Panevová (1986) *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, D. Reidel, Dordrecht, Holandia.

S. Sharoff (2004) *Methods and tools for development of the Russian Reference Corpus*, w: A. Wilson, D. Archer, P. Rayson (red.), *Corpus Linguistics Around the World*, t. 56 serii *Language and Computers. Studies in Practical Linguistics*, s. 167–180, Rodopi, Amsterdam, Holandia.

W. Skut, T. Brants, B. Krenn, H. Uszkoreit (1998) *A Linguistically Interpreted Corpus of German Newspaper Text*, w: *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, s. 705–711, Saarbrücken, Niemcy.

W. Skut, B. Krenn, T. Brants, H. Uszkoreit (1997) *An Annotation Scheme for Free Word Order Languages*, w: *Proceedings of the 5th ACL Conference on Applied Natural Language Processing (ANLP-97)*, s. 88–96, Washington, DC.

P. Smrž (2004) *Qaulity Control for Wordnet Development*, w: P. Sojka, K. Pala, P. Smrž, C. Fellbaum, P. Vossen (red.), *Proceedings of the 2nd International WordNet Conference (GWC 2004)*, s. 206–212, Masaryk University, Brno, Czechy.

P. Sojka, V. Matoušek, P. Mautner, K. Pala, I. Kopeček (red.) (1998) *Proceedings of the 1st International Workshop on Text, Speech and Dialogue*, Masaryk University, Brno, Czechy.

M. Spevack (1968–70) *Complete and Systematic Concordance to the works of Shakespeare*, G. Olms, Hildesheim, Niemcy.

— (1972) *Shakespeare English: The Core Vocabulary*, RNL, t. 3, nr ii, s. 106–122.

J. Stein (red.) (1967) *The Random House Dictionary of the English Language*, Random House, Nowy Jork, NY.

M. Szupryczyńska (1973) *Syntaktyczna klasyfikacja czasowników przybiernikowych*, Państwowe Wydawnictwo Naukowe, Poznań.

A. Taylor, M. P. Marcus, B. Santorini (2003) *The Penn Ttrebank: An Overview*, w: Abeillé (2003), s. 5–22.

TEI P5 (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Internet.

W. Teubert, R. Krishnamurthy (red.) (2007) *Corpus Linguistics*, Critical Concepts in Linguistics, Routlege, Abington, Wielka Brytania; Nowy Jork, NY.

M. Woliński (2006) *Morfeusz — a Practical Tool for the Morphological Analysis of Polish*, w: M. A. Kłopotek, S. T. Wierzchoń, K. Trojanowski (red.), *Proceedings of the Intelligent Information Systems New Trends in Intelligent Information Processing and Web Mining IIS:IIPWM'06*, Advances in Soft Computing, s. 503–512, Springer-Verlag, Ustroń.

A. Zalizniak (1977) *Grammaticzeskij slovar' russkogo jazyka*, Russkij Jazyk, Moskwa, Rosja.

H. Zinsmeister, J. Kuhn, S. Dipper (2001a) *From LFG Structures to TIGER Treebank Annotations*, w: *Proceedings of the Third Workshop on Linguistically Interpreted Corpora (LINC 2001)*, Leuven, Belgia.

H. Zinsmeister, J. Kuhn, B. Schrader, S. Dipper (2001b) *TIGER Transfer — From LFG Structures to the TIGER Treebank*, Rap. tech., Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

G. K. Zipf (1935) *The Psycho-Biology of Language: An Introduction to Dynamic Philology*, Houghton Mifflin, Boston, MA.

Pracę zgłosił Adam Przepiórkowski

Adres autorki: Elżbieta Hajnicz
Instytut Podstaw Informatyki PAN
ul. Ordona 21
01-237 Warszawa
Polska
e-mail: Elzbieta.Hajnicz@ipipan.waw.pl

Symbol klasyfikacji rzeczowej: CR: I.2.7

Nakład 100 egzemplarzy. Papier kserograficzny klasy III. Oddano do druku w grudniu 2011. Wydawnictwo IPI PAN. ISSN 0138-0648.