NARODOWY KORPUS JĘZYKA POLSKIEGO

Praca zbiorowa pod redakcją

Adama Przepiórkowskiego

Mirosława Bańko

Rafała L. Górskiego

Barbary Lewandowskiej-Tomaszczyk

NARODOWY KORPUS JĘZYKA POLSKIEGO

WYDAWNICTWO NAUKOWE PWN
WARSZAWA 2012

Projekt okładki i stron tytułowych
Przemysław Spiechowski

Wydawca
Magdalena Ścibor

Redaktor
Joanna Cierkońska

Produkcja
Edyta Kunowska

Warszawa 2012

ISBN 978-83-01-16700-4

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

Spis treści

Część I

Wstęp

1

Narodowy Korpus Języka Polskiego: geneza i dzień dzisiejszy

Barbara Lewandowska-Tomaszczyk, Mirosław Bańko, Rafał L. Górski, Marek Łaziński, Piotr Pęzik, Adam Przepiórkowski

Nareszcie mamy narodowy korpus polszczyzny – dostępny publicznie i w dodatku bezpłatnie. Użytkownicy, którzy zechcą wykorzystać go do swoich celów, mogą uznać, że jest niedostatecznie zróżnicowany, nie dość precyzyjnie oznakowany – te i inne zarzuty łatwo sobie wyobrazić. Łatwo je także odparować: praca takich rozmiarów jest wynikiem różnych kompromisów, poza tym trudno jest wszystkich zadowolić.

Być może niektórzy uznają, że NKJP się niewłaściwie nazywa, gdyż słowo *narodowy* ma w polskim uchu konotacje patriotyczne i w nazwie projektu naukowego brzmi pretensjonalnie. Wątpliwości związane z nazwą twórcy korpusu sami mieli na początku prac, ale wydaje się, że niesłusznie; słowo *narodowy* w nazwie NKJP kontynuuje tradycje polskiej leksykografii. Przypomnijmy, że Linde swój słownik nazywał narodowym (zob. Matuszczyk 2006: 99–104) i że pod nagłówkiem „słowniki narodowe" zgrupował podobne dzieła Grzegorczyk (1967) w swojej bibliografii. Także w leksykografii encyklopedycznej funkcjonuje kategoria encyklopedii narodowych (np. Olkiewicz 1988). Co więcej, jak wiele innych narodów mamy Bibliotekę Narodową, której zbiory z upływem czasu będą coraz bardziej dygitalizowane, co w końcu upodobni ją do korpusu, a nawet uczyni korpusem *in potentia*.

1.1. Krótki rys historyczny

Od lat sześćdziesiątych dwudziestego wieku wzrastało zainteresowanie językoznawców badaniem częstotliwości użycia różnych form językowych oraz prawdopodobieństwa ich występowania w różnych odmianach języka. Zainteresowanie

to zbiegło się w czasie z dynamicznym rozwojem komputerów i nowych technik informacyjnych. Komputer pozwolił językoznawcom na magazynowanie i szybką obróbkę dużej ilości danych językowych. Nowo powstała dyscyplina lingwistyczna, *językoznawstwo korpusowe*, rozwijająca się obecnie niezwykle dynamicznie w wielu krajach świata, pozwoliła na badania języka na szeroka skalę. Zbiory językowe, gromadzone w *korpusach językowych* czyli komputerowych zbiorach autentycznych tekstów językowych, mówionych i pisanych, reprezentują różne odmiany, style i typy tekstu.

Pierwszy powszechnie używany językowy korpus komputerowy to korpus amerykańskiej odmiany języka angielskiego, zawierający zaledwie jeden milion wyrazów, ale bezprecedensowy wtedy, zebrany w latach sześćdziesiątych ubiegłego wieku przez Henry'ego Kučerę i Nelsona Francisa (Kučera i Francis 1967). Obecnie zbiory językowe sięgają setek milionów, a nawet miliardów, jednostek. Autentyczne materiały korpusowe używane są dziś do różnych zadań, takich jak opracowywanie słowników i materiałów dydaktycznych, przekładu, jak również do studiów literackich, kulturowych i in. Komputery umożliwiają magazynowanie i analizowanie dużo większych zbiorów materiału językowego niż tradycyjne metody językoznawcze. Ponadto, co jest szczególnie ważne, pozwalają na weryfikowanie intuicji językowych, pokazując użycie i frekwencję zarówno form krótszych, jak i wzorców współwystępowania różnych form językowych i ich częstotliwości użycia. Pozwalają także na szeroko zakrojone badania języka różnych grup użytkowników, jak i kwantyfikację wyników badań dialektologicznych czy socjolingwistycznych na dużą skalę w zależności od typu dyskursu, stylu, czy indywidualnych preferencji użytkownika języka. Zbiory językowe w korpusach są także niezbędne do zastosowań w szerszej gamie działań językoznawstwa komputerowego, np. dla celów przetwarzania języka naturalnego z wykorzystaniem metod uczenia maszynowego.

Najliczniejsze i najbardziej różnorodne zbiory zawierają narodowe korpusy angielskie – brytyjskie (British National Corpus, Bank of English i in.) i amerykańskie (Corpus of Contemporary American English, American National Corpus, Google Books: American English i in.). Zbiory innych języków rozwijają się dynamicznie. Są to m.in. chiński korpus mandaryńskiego, korpusy języka niderlandzkiego i duńskiego, estoński korpus języka prawa, korpusy odmian języka francuskiego, zbiory tekstów niektórych języków celtyckich (język irlandzki oraz język mański), włoski korpus CORIS/CODIS, teksty norweskie *Oslo Corpus of Tagged Norwegian Texts*, korpus brazylijskiej odmiany języka portugalskiego, „bank językowy" tekstów szwedzkich, korpusy europejskiej i południowo-amerykańskich odmian języka hiszpańskiego, korpus języka tagalog, korpus literatury malajskiej i korpusy innych języków południowo-wschodniej Azji.

Od kilku lat szczególnie dynamicznie rozwijają się korpusy narodowe innych języków słowiańskich. Od roku 1994, kiedy rozpoczęły się prace nad Czeskim Korpusem Narodowym, powstały wielkie referencyjne korpusy większości języków słowiańskich, z których chorwacki, czeski, polski, rosyjski, słowacki i słoweński dostępne są w Internecie z wyszukiwarkami morfologicznymi (trzeba podkreślić, że morfologia słowiańska stanowi dla technologii większe wyzwanie niż romańska czy germańska). Narodowy Korpus Języka Poskiego współorganizował konferencję założycielską Komisji Korpusowej Międzynarodowego Komitetu Slawistów (por. Slavicorp 2012).

Najbardziej naturalnym użyciem korpusów językowych było zawsze zastosowanie autentycznych materiałów językowych w pracach słownikowych. Pierwszy słownik języka polskiego oparty na korpusach w nowoczesnym rozumieniu tego pojęcia to listy frekwencyjne różnych odmian polszczyzny z lat sześćdziesiątych dwudziestego wieku (Kurcz i in. 1974a, b, 1976, 1977, Lewicki i in. 1975). Dane te posłużyły także do opracowania nowego polskiego słownika frekwencyjnego (Kurcz i in. 1990).

Budowa narzędzi korpusowych jest równoległym do zbierania danych zestawem działań językoznawstwa korpusowego, niezbędnym do efektywnego używania korpusów. Języki fleksyjne, takie jak język polski, stanowią spore wyzwanie dla automatycznej analizy morfologicznej i składniowej. W języku polskim częsta jest wieloznaczność form gramatycznych, zarówno między częściami mowy, jak również w kręgu kategorii gramatycznych, takich jak liczba i przypadek. Szczególnie istotne dla automatycznego rozpoznawania form wyrazowych są więc propozycje autorów analizatorów morfologicznych języka polskiego, programów ujednoznaczniających oraz testów do ich weryfikacji i oceny.

W Polsce aktywnie działało od lat dziewięćdziesiątych kilka grup językoznawczych, informatycznych i leksykograficznych, które zajmowały się zarówno zbieraniem danych korpusowych, jak i tworzeniem narzędzi do ich opracowywania, m.in. zespoły w Instytucie Podstaw Informatyki Polskiej Akademii Nauk (`http://nlp.ipipan.waw.pl/`), zespół w Instytucie Języka Polskiego Polskiej Akademii Nauk w Krakowie (`http://www.ijp-pan.krakow.pl/`), zespół korpusowy PELCRA (`http://pelcra.pl/`) w Katedrze Języka Angielskiego i Językoznawstwa Stosowanego w Uniwersytecie Łódzkim oraz zespół korpusowy w Redakcji Słowników Języka Polskiego Wydawnictwa Naukowego PWN (`http://korpus.pwn.pl/`).

IPI PAN Zespół Inżynierii Lingwistycznej został założony – i był przez wiele lat kierowany – przez prof. Leonarda Bolca. W latach dziewięćdziesiątych ubiegłego wieku działalność zespołu dotyczyła głównie przetwarzania składniowego języka polskiego, przede wszystkim w ramach teorii formalnej Head-driven Phrase

Structure Grammar. Ukoronowaniem tej działalności była wydana w 2002 roku monografia *Formalny opis języka polskiego: teoria i implementacja* (Przepiórkowski i in. 2002).

Zespół rozpoczął badania korpusowe na początku obecnego stulecia. Ponieważ jednak w tym czasie nie był publicznie dostępny duży i odpowiednio znakowany korpus języka polskiego, w lipcu 2000 roku został złożony do Komitetu Badań Naukowych wniosek o finansowanie budowy Korpusu IPI PAN. Realizacja projektu trwała od kwietnia 2001 do marca 2004 roku i zaowocowała stworzeniem ówcześnie największego – choć nie zrównoważonego ani nie reprezentatywnego w żadnym sensie – korpusu języka polskiego. Korpus ten liczył 250 milionów segmentów (ponad 200 milionów tradycyjnie rozumianych słów) i był pierwszym dużym polskim korpusem znakowanym morfosyntaktycznie (`http://korpus.pl/`).

Obecnie prace zespołu koncentrują się wokół kilku projektów europejskich: CLARIN, CESAR (w ramach konsorcjum META-NET) i ATLAS, oraz krajowych: NEKST (*Adaptacyjny system wspomagający rozwiązywanie problemów w oparciu o analizę treści dostępnych źródeł elektronicznych*), *Budowa banku drzew składniowych dla języka polskiego z wykorzystaniem automatycznej analizy składniowej* i *Komputerowe metody identyfikacji nawiązań w tekstach polskich*, a także – w mniejszym stopniu – SYNAT (*Utworzenie uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla sieciowych zasobów wiedzy dla nauki, edukacji i otwartego społeczeństwa wiedzy*).

IJP PAN Początki prac nad korpusem IJP PAN ściśle wiążą się z zarzuconym później pomysłem słownika języka polskiego. Korpus miał być dla niego bazą empiryczną. Gdy w początkach lat dziewięćdziesiątych okazało się, że w ówczesnej sytuacji projekt nie może dojść do skutku, korpus zaczęto wykorzystywać do badań prowadzonych w Instytucie, w szczególności stanowił podstawowe źródło dla badań nad semantyką i składnią czasowników polskich, projektu badawczego, który ma zaowocować słownikiem. Był też przez cały czas rozbudowywany, nie planowano natomiast publicznego udostępnienia tego korpusu, jakkolwiek – szczególnie zanim powstał korpus PWN – korzystało z niego wielu naukowców spoza IJP. Na marginesie warto dodać, że w IJP PAN został stworzony również korpus staropolski, obejmujący wszystkie ciągłe teksty do roku 1500. Korpus ten jest publicznie dostępny na płytach CD (Twardzik 2006). To jak dotąd jedyny korpus historyczny języka polskiego.

PELCRA UŁ Jedną z najstarszych grup językoznawstwa korpusowego jest zespół PELCRA, działający w Katedrze Języka Angielskiego Uniwersytetu Łódzkiego od 1995 roku, początkowo we współpracy ze znanym językoznawcą

korpusowym Anthonym McEnerym i pracownikami Katedry Językoznawstwa i Współczesnego Języka Angielskiego w Uniwersytecie w Lancaster w Wielkiej Brytanii. Były to też lata, w których powstawał największy znany korpus językowy British National Corpus, czyli Brytyjski Korpus Narodowy. Kontakty te zaowocowały wspólnymi badaniami rozpoczętymi w ramach projektu, który został nazwany PELCRA (*Polish and English Language Corpora for Research and Application*). Obecne zasoby zespołu PELCRA to blisko stumilionowy korpus PELCRA, reprezentujący proporcjonalnie różne typy i odmiany tekstów, zarówno języka mówionego jak i pisanego, dostępny w witrynie `http://korpus.ia.uni.lodz.pl/`, z wyszukiwarką, której współautorem jest Piotr Pęzik, oraz korpus tekstów mówionych liczący ponad 600 tysięcy słów.

Polski Korpus Uczniowski Języka Angielskiego w zasobach PELCRA, który liczy obecnie 1,5 miliona słów, docelowo zaś – 3 miliony, zawiera angielskie dane językowe Polaków uczących się języka angielskiego (grant MNiSW nr N N104 205039, kierowany przez Piotra Pęzika). W toku są prace nad budową korpusów paralelnych i porównywalnych angielsko-polskich i polsko-angielskich (europejski projekt CESAR / META-NET, numer ICT-PSP 271022).

Prace badawcze członków zespołu PELCRA koncentrują się na kilku kręgach tematycznych. Należy do nich analiza i opis języka angielskiego, także w aspekcie kontrastywnym w porównaniu z językiem polskim oraz języka polskiego – na tle różnic i podobieństw z językiem angielskim. Ponadto znajdują się wśród nich badania nad rozwijaniem technik i materiałów do nauczania języków obcych, analizy przekładów oraz studia nad problematyką interkulturowości.

PWN Zespół korpusowy utworzono w Redakcji Słowników Języka Polskiego w roku 1997 w trakcie prac nad pierwszym słownikiem języka polskiego opartym na korpusie – *Innym słownikiem języka polskiego* (Bańko 2000) oraz bardziej tradycyjnym *Uniwersalnym słownikiem języka polskiego* (Dubisz 2003). Personalnie zespół Korpusu Języka Polskiego PWN i Redakcji Słowników Języka Polskiego był w naturalny sposób związany z Wydziałem Polonistyki Uniwersytetu Warszawskiego. Korpus PWN miał być podstawą opisu leksykograficznego i bazą różnorodnych przykładów ilustrujących znaczenie haseł w słownikach. Dlatego największy nacisk położono w nim na różnorodność tematyczną i stylistyczną tekstów, a także na ich reprezentatywność z punktu widzenia polskiej tradycji literackiej (w skład korpusu weszła m.in. cała lista lektur z literatury polskiej na poziomie programu maturalnego). Korpus PWN był także wykorzystywany przez językoznawców spoza wydawnictwa. Zrównoważony korpus z czasem osiągnął nieco ponad 100 milionów słów, a 40-milionowa próbka dostępna jest w Internecie z prostą wyszukiwarką: `http://korpus.pwn.pl/`.

1.2. Narodowy Korpus Języka Polskiego

Mimo opisanych powyżej sukcesów sytuacji nie można było uznać za zadowalającą. Wciąż bowiem nie dysponowaliśmy korpusem, który byłby równocześnie duży, zróżnicowany, reprezentatywny i anotowany morfosyntaktycznie. Realizowane projekty stworzyły narzędzia badawcze, które spełniały tylko niektóre z powyższych postulatów. Równocześnie narastało przekonanie, że dotychczasowe rozproszenie wysiłków jest niekorzystne[1]. W roku 2006, z inspiracji Prezydium Komitetu Językoznawstwa Polskiej Akademii Nauk i przy poparciu jego ówczesnego przewodniczącego prof. dr. hab. Stanisława Gajdy, zawiązało się konsorcjum *Narodowy Korpus Języka Polskiego* (`http://nkjp.pl/`), w skład którego weszły trzy ośrodki akademickie: IPI PAN, reprezentowany przez Adama Przepiórkowskiego, IJP PAN (Rafał Górski) i UŁ (Barbara Lewandowska-Tomaszczyk), oraz wydawnictwo PWN (Mirosław Bańko i Marek Łaziński). W następnym roku zespół NKJP uzyskał grant rozwojowy (nr R17 003 03, od grudnia 2007 do grudnia 2010, przedłużony do czerwca 2011 roku), którego głównym celem była budowa NKJP oraz opracowanie narzędzi do jego wykorzystywania. Za zgodą wszystkich członków nowo powstałego zespołu koordynatorem projektu został IPI PAN, jego kierownikiem zaś dr hab. Adam Przepiórkowski.

Znaczące części korpusów partnerów projektu weszły w skład Narodowego Korpusu Języka Polskiego. IPI PAN przekazało NKJP cały zebrany wcześniej korpus, liczący ponad 200 milion słów. Z zasobów Korpusu Języka Polskiego PWN do NKJP weszło na początku projektu 50 milionów słów, kolejne 200 milionów tekstów książkowych, w tym w dużej części literackich oraz prasowych, zebrano w trakcie realizacji projektu. Z zasobów korpusu PELCRA do Narodowego Korpusu Języka Polskiego także weszło kilkadziesiąt milionów słów, dalsze ponad 660 milionów zostało zebranych w pracach projektowych. Ponad 2 miliony słów stanowią nowe dane mówione (w skład korpusu zostaje włączonych 1,9 miliona słów danych konwersacyjnych, które NKJP zamierza udostępnić na licencji niekomercyjnej). Najnowsze materiały zespołu PELCRA to kilkusetmilionowe zbiory tekstów internetowych, których istotna część zasiliła zasoby NKJP.

W ramach projektu powstały i były rozwijane różne formy narzędzi korpusowych. Piotr Pęzik z zespołu PELCRA jest autorem kilku aplikacji korpusowych dla NKJP: wyszukiwarki korpusowej do danych NKJP, wyszukiwarki do danych

[1] Z pewnością trudno to początkowe rozproszenie wysiłków uznać za zjawisko pozytywne. Należy na to popatrzeć także z drugiej strony. NKJP tworzył zespół badaczy, którzy po pierwsze, mieli już niemałe doświadczenie w zakresie opracowywania korpusów, a po drugie intensywnie używali ich w swoich badaniach, zdawali więc sobie sprawę z tego, czego od korpusu oczekuje jego użytkownik.

mówionych NKJP (`http://nkjp.uni.lodz.pl/spoken.jsp`) oraz dla zainteresowanych wyszukiwarki SlopeQ do BNC. Jest też autorem kolokatora do NKJP oraz Automatycznego Słownika Kolokacji (ASK) (dostępna wersja demo). Zmodyfikowana została także przeszukiwarka Poliqarp, służąca wcześniej do obsługi Korpusu IPI PAN. Stworzono także liczne narzędzia do przetwarzania tekstów, szczegółowo opisane w dalszych rozdziałach; są to m.in.: tager morfosyntaktyczny PANTERA, nowa gramatyka powierzchniowa języka polskiego dla parsera Spejd, narzędzia identyfikujące w tekstach jednostki nazewnicze oraz prototypowy system ujednoznaczniania sensów słów.

Dziś, w roku 2011, możemy stwierdzić, że Narodowy Korpus Języka Polskiego, największy, morfologicznie anotowany zbiór danych języka polskiego, jest faktem. NKJP jest dostępny bezpłatnie dla wszystkich chętnych i zainteresowanych, tom zaś, który Państwu przedstawiamy to pokłosie prawie czteroletnich wysiłków zespołu NKJP.

Z korpusu korzysta już bardzo wielu użykowników. Danymi NKJP posługuje się przede wszystkim zespół powstającego *Wielkiego słownika języka polskiego* (Żmigrodzki i in. 2007), jest on również wykorzystywany w badaniach kontrastywnych z udziałem języka polskiego (np. Dziwirek i Lewandowska-Tomaszczyk 2010), w praktyce i dydaktyce translatoryki (Pęzik 2011), a także w Poradni Językowej PWN. Wydawnictwa UW zapowiadają już nawet pierwszy słownik powstały na podstawie NKJP – nosi on tytuł *Ludzie i miejsca w języku*, a gromadzi frazy odimienne typu *puszka Pandory* lub *jajko Kolumba* (autorami są Maciej Czeszewski i Katarzyna Foremniak).

Nie jest to naturalnie koniec naszych prac. Czekają nas dalsze wyzwania. Aby pozostał największym narzędziem referencyjnym dla rzeszy użytkowników, NKJP musi być ustawicznie uzupełniany, ulepszany, nadzorowany i monitorowany. Ponadto konstrukcja zarówno programów automatycznego wydobywania znaczeń z danych korpusowych, jak i programów ujednoznaczniających to problematyka, z którą zmaga się duża część językoznawstwa informatycznego. Naturalne jest także, że dla rozpowszechniania zarówno samych danych NKJP, jak i narzędzi do ich wykorzystywania konieczne są wersje różnych aplikacji, mogących zdalnie obsługiwać jednocześnie wielu użytkowników, o różnych potrzebach poznawczych i celach aplikacyjnych.

1.3. Podziękowania

Prace opisane w niniejszej publikacji były finansowane ze środków na naukę w latach 2007–2011 w ramach projektu rozwojowego „Narodowy Korpus Języka Polskiego” (nr R17 003 03). Spis licznych dobroczyńców NKJP – przede wszystkim

wydawnictw i autorów, ale także innych grup badawczych i zespołów projektów, z którymi NKJP współpracował – oraz wykonawców niniejszego projektu, znajduje się na stronach `http://nkjp.pl/` (w zakładkach PODZIĘKOWANIA, TEKSTY KORPUSU i ZESPÓŁ). Wszystkim tym osobom i instytucjom oraz rzeszy innych, którzy pomagali nam w pracach, w tym miejscu raz jeszcze dziękujemy.

Część II

Struktura korpusu

2

Typologia tekstów w NKJP

Rafał L. Górski, Marek Łaziński

Ustalenie właściwej typologii tekstów ma dla każdego korpusu podwójne znacznie. Przede wszystkim pozwala użytkownikowi korpusu ograniczać przeszukiwanie do wybranych typów tekstu. Zwiększa więc tym samym funkcjonalność korpusu. Właściwa typologia tekstów musi także poprzedzać decyzję co do ustalenia budowy korpusu.

Każda decyzja dotycząca typologii tekstów jest w pewnym stopniu decyzją arbitralną, choćby dlatego, że wśród specjalistów w tym zakresie brak pełnej zgody. Przy tworzeniu typologii tekstów staraliśmy się oprzeć na istniejących opracowaniach i typologiach wypracowanych na gruncie językoznawstwa polskiego. Są to przede wszystkim ustalenia, które wypracowali Klemensiewicz (1982) i Gajda (1995).

Trzeba jednak wziąć pod uwagę, że większość typologii, które nie zostały wypracowane do celów praktycznych, nie może być wprost zastosowana do naszych celów, ponieważ – co oczywiste – językoznawcy skupiają się na najistotniejszych typach tekstów, podczas gdy marginalne – bądź to dlatego, że rzadsze, bądź dlatego, że kulturowo mniej istotne – pozostają poza obrębem ich zainteresowania.

Choć więc kategorie takie jak powieść czy krótka wiadomość prasowa są dość dobrze zdefiniowane i niekontrowersyjne, to jednak istnieje szereg tekstów, które nie mieszczą się w ramach istniejących typologii. I tak np. niewiele uwagi poświęca się *instrukcji obsługi*, a jeszcze mniej dość istotnemu typowi tekstu, jakim jest *przewodnik turystyczny*[1].

Ponadto w nauce ugruntowana jest opinia, że w stylistyce można mówić o pewnych idealnych wykładnikach stylu, tymczasem poszczególne teksty albo charakteryzują się tylko niektórymi z tych wykładników, albo też wykazują przynależność do dwu różnych typów tekstu. Innymi słowy tekst może równocześnie przynależeć np. do publicystyki i literatury pięknej.

[1] Więcej o potrzebie rozszerzenia tradycyjnej genologii por. Wojtak 2008.

Mając na uwadze powyższe uwarunkowania, twórcy korpusu starają się stworzyć listę typów tekstów, która byłaby na tyle wyczerpująca, by każdy tekst dało się przypisać do danego typu i zdefiniować te kategorie tak, żeby dało się jednoznacznie przypisać do któregoś typu teksty z pogranicza. Tym samym tego rodzaju definicje mają cel praktyczny i mogą np. zwracać uwagę na szczegóły drugorzędne, o ile tylko stanowią one wygodne kryterium rozstrzygnięć.

W wypadku nowo powstającego korpusu stworzenie tego rodzaju typologii jest o tyle skomplikowane, że łatwo nie przewidzieć istnienia pewnych marginalnych kategorii. Zespół NKJP miał łatwiejszą sytuację, ponieważ robocze wersje typologii można było przetestować na zasobach scalanych korpusów.

2.1. Przegląd istniejących typologii

Nie istnieje uniwersalna typologia tekstów przyjmowana w językoznawstwie korpusowym. Dlatego też nie zawsze daje się porównać ze sobą dwa różne korpusy. Generalnie przyjmuje się fundamentalny podział na teksty fikcjonalne i informacyjne (ang. *imaginative* oraz *informative*)[2], te drugie dzieli się na publicystykę i rozmaicie etykietkowane „pozostałe teksty informacyjne". Głębsze podziały mają jednak zazwyczaj charakter tematyczny (Čermák 2001, Kučera 2002). Tylko tak bowiem można rozumieć kategorie „handel i finanse" (ang. *commerce & finance*) czy „czas wolny" (ang. *leisure*). Jak widać, najwyższe piętra takich podziałów mają charakter stylistyczny (funkcjonalny), natomiast kategorie bardziej szczegółowe – tematyczny[3].

Większość korpusów uwzględnia podział tekstów według ich kanału. British National Corpus uwzględnia także kategorię „poziomu". Dla tekstów literackich kategoria ta ma odzwierciedlać ambicje artystyczne, a dla tekstów informacyjnych – stopień ich hermetyczności.

2.2. Typologia tekstów w NKJP

Zakładamy, że każdy z tekstów korpusu zalicza się do jednego z typów wymienionych poniżej:

[2] Trzeba wyraźnie zaznaczyć, że podział na teksty fikcjonalne i niefikcjonalne w większości korpusów językowych jest sprzeczny z intuicyjnym wyróżnieniem tzw. literatury pięknej. Wiele książek nagrodzonych w konkursach literackich i omawianych na zajęciach uniwersyteckich z historii literatury klasyfikuje się w korpusach jako non-fiction. Także NKJP traktuje jako non-fiction nie tylko reportaże Ryszarda Kapuścińskiego, ale też klasyczne eseje Jarosława Marka Rymkiewicza.

[3] Użyteczną, choć bardzo pobieżną historię typologii tekstów w różnych korpusach angielskich przedstawia Holmes-Higgin i in. (1994).

1. literatura piękna
   a) proza,
   b) poezja,
   c) dramat,
2. literatura faktu,
3. publicystyka i krótkie wiadomości prasowe,
4. typ naukowo-dydaktyczny,
5. typ informacyjno-poradnikowy,
6. książka niebeletrystyczna niesklasyfikowana,
7. inne teksty pisane
   a) typ urzędowo-kancelaryjny,
   b) teksty perswazyjne (ogłoszenia, reklamy, propaganda polityczna),
   c) krótkie teksty instruktażowe ,
8. listy,
9. Internet
   a) interaktywne strony WWW (fora, chaty, listy dyskusyjne itp.),
   b) statyczne strony WWW,
10. teksty mówione konwersacyjne,
11. teksty mówione medialne,
12. teksty quasi-mówione.

Podział na typy to przecięcie tradycyjnych podziałów stylistycznych i gatunkowych. Wynika on z inherentnych cech struktury tekstu – stylistycznych lub gatunkowych, czasem – lecz nie często – implikowanych przez poruszaną tematykę, lecz zasadniczo od podziału tematycznego niezależnych. Podział na typy jest często zintegrowany z podziałem na kanały przekazu, ale z definicji te dwa podziały pozostają odrębne.

Na początek zauważmy, że powyższa lista zakłada pewną hierarchizację. I tak literatura piękna dzieli się dalej na: prozę, poezję i dramat (wierszem lub prozą). Dla ścisłości dodajmy jednak, że nie wszystkie teksty były oznaczane z dokładnością do podtypu. Generalnie jedynie w wypadku poezji i (szeroko rozumianego) dramatu zawsze staraliśmy się uszczegółowić klasyfikację. Tak więc można domyślnie przyjąć, że kategoria „literatura piękna" (bez dalszych dookreśleń) to proza artystyczna.

Nieco inaczej ma się rzecz z typem 7: inne teksty pisane. Typy a), b) i c) są jego podtypami. Ponieważ jest to typ niezwykle zróżnicowany, nie ma ani możliwości, ani potrzeby wprowadzania wyczerpującego podziału szczegółowego. Wyczerpującego, to znaczy takiego zestawu podtypów, który obejmowałby wszystkie istniejące teksty pisane. Taka klasyfikacja – jeśli naprawdę miałaby być wyczerpująca – musiałaby być niezwykle szczegółowa; równocześnie każdy

z tych typów byłby reprezentowany w korpusie przez tak niewielką liczbę słów, że użytkownik nie odnosiłby żadnej korzyści z tego rodzaju podziałów.

Również typ 3 jest w istocie rzeczy heterogeniczny. Publicystyka bowiem różni się dość istotnie od tzw. krótkich informacji prasowych. Połączenie tych dwu kategorii w jedną jest podyktowane względem praktycznym – często pochodzą z tych samych tytułów prasowych i niełatwo je rozdzielić.

Większość korpusów wydziela jako odrębną kategorię teksty mówione. My postąpiliśmy wbrew tej tendencji, w założonej typologii nie ma bowiem takiej kategorii. Zamiast niej zaproponowaliśmy trzy niezależne typy: teksty mówione konwersacyjne, mówione medialne i quasi-mówione. Uznaliśmy mianowicie, że są to na tyle różne typy tekstów, że nie należy ich łączyć. Jeśli zainteresowany użytkownik uzna potrzebę badania ich łącznie, zawsze może to zrobić, ustawiając odpowiednio warunki przeszukiwania.

2.3. Kanał

Tekst charakteryzowany jest – oprócz typu – także przez kanał, a więc sposób, w jaki technicznie rzecz biorąc, zostaje przekazany od nadawcy do odbiorcy. Wyróżniamy następujące kanały:

1. prasa
   a) prasa – dziennik,
   b) prasa – tygodnik,
   c) prasa – miesięcznik,
   d) prasa – inne,
2. książka,
3. Internet,
4. mówiony,
5. ulotka,
6. rękopis.

Zasadniczo kanały te są intuicyjnie łatwo uchwytne, ale konieczne jest tu kilka słów komentarza. Za kanał uznawaliśmy to, co jest dla danego tekstu prymarne. Większość tekstów prasowych w istocie pozyskano z Internetu albo przynajmniej z wersji elektronicznych przygotowanych do publikacji w Internecie. Nie zmienia to jednak faktu, że kanałem, jakim tekst prasowy, ma dotrzeć do czytelnika jest wydanie papierowe. Wersja internetowa jest z zasady jedynie kopią wersji papierowej, nie zawsze zresztą dokładną[4]. Jeśli jednak czasopismo ukazywało się jedynie w Internecie, oczywiście przypisywano mu taki właśnie kanał.

[4] Trzeba sobie jednak zdawać sprawę z tego, że sytuacja ta gwałtownie się zmienia, gdyż coraz więcej osób czyta prasę wyłącznie w Internecie, a portale największych gazet oprócz przekazania wiadomości proponują usługi i serwisy charakterystyczne dla klasycznych portali internetowych.

Uznanie częstotliwości pojawiania się danego tytułu prasowego za kryterium przydzielające do kanału też może być dyskusyjne. Decyzja ta była podyktowana względami ekonomii: chodziło o to, by nie tracić informacji o częstotliwości wydawania czasopisma, skądinąd istotnej zarówno dla twórców korpusu jak i jego użytkowników, a równocześnie nie ustanawiać osobnego znacznika tej informacji.

W nielicznych przypadkach nasuwały się pewne wątpliwości, np. czy kanałem jest książka, czy prasa. W takiej sytuacji rozstrzygające było to, czy dana jednostka bibliograficzna jest zarejestrowana w systemie ISSN czy też ISBN[5]. Z podobnego powodu (oznaczenie ISSN) uznaliśmy, że kanałem protokołu parlamentarnego jest prasa. Należy bowiem zwrócić uwagę, że jeśli wydawca opatruje jednostkę bibliograficzną numerem ISSN, to jest to jego deklaracja co do charakteru publikacji, który my nazywamy kanałem.

W tym miejscu warto zauważyć, że istnieje jedynie luźny związek między kanałem a typem tekstu. Ściśle rzecz biorąc, pewne kanały służą prymarnie do transmisji określonych typów tekstu. I tak np. prasa jest nośnikiem publicystyki i krótkich informacji prasowych. Z kolei publicystyka jest ogłaszana również w książkach i Internecie; istnieją też czasopisma publikujące wyłącznie teksty naukowe bądź też literackie.

Tak więc kanał przekazu tekstu był istotną wskazówką, pozwalającą przyporządkować tekst do danego typu, ale w żadnym razie nie rozstrzygającą. Tym bardziej należy się liczyć z tym, że w korpusie pojawią się teksty o nietypowym zestawieniu kanału i typu.

2.4. Klasyfikacja tematyczna

Teksty korpusu są również klasyfikowane pod względem tematycznym. W tym wypadku przyjmujemy bez zmian dwie klasyfikacje Biblioteki Narodowej, a mianowicie kod Uniwersalnej Klasyfikacji Dziesiętnej i system haseł tematycznych. Służą one do kontroli zróżnicowania tematycznego korpusu, a także w niektórych wypadkach stanowią cenną wskazówkę co do przynależności tekstu do danego typu.

Należy zaznaczyć, że dane te są zawarte w nagłówkach, ale nie w metadanych wyświetlanych w programach konkordancyjnych, nie są więc dostępne dla zewnętrznego użytkownika korpusu.

[5] Dla ścisłości należy dodać, że dana jednostka może być oznaczona równocześnie ISSN i ISBN, a mianowicie wtedy, gdy jest to książka ukazująca się w serii; w takiej sytuacji decydujemy się na kanał książka.

2.5. Definiowanie typów tekstów

Opisana powyżej hierarchia znajduje zastosowanie w kryteriach przypisywania konkretnych tekstów do poszczególnych typów. Jak wspomnieliśmy wyżej, na potrzeby tworzenia korpusu trzeba opracować szczegółową procedurę tej czynności, taką, która na tyle, na ile jest to możliwe, daje powtarzalne i niearbitralne wyniki. Podkreślamy: na ile to możliwe, ponieważ pewien margines niepewności zawsze pozostaje.

Pierwszy krok to ustalenie kanału tekstu. Zasadniczo jeśli kanałem jest prasa, przypisujemy tekst do typu „publicystyka i krótkie wiadomości prasowe". Wyjątek stanowią czasopisma naukowe (np. *Studia ad Bibliothecarum Scientiam Pertinentia*) czy szerzej – specjalistyczne oraz literackie (np. *Esensja*). Można więc powiedzieć, że typ ten definiowany jest negatywnie – każdy tekst drukowany w prasie należy do omawianego typu, o ile na podstawie innych (niżej wymienionych) kryteriów nie przynależy do innego. Inaczej ma się rzecz z sytuacją odwrotną (kanał – książka, typ tekstu – publicystyka i krótkie wiadomości prasowe), czyli tak zwaną książką publicystyczną. W tym wypadku kierowaliśmy się następującymi kryteriami: tekst zawiera raczej przekaz opinii niż fakty czy narrację; jego zadaniem jest raczej wywieranie wpływu niż czysta informacja. Dodatkowo potwierdzeniem decyzji był fakt, że hasło klasyfikacji Biblioteki Narodowej zawierało element *publicystyka*, np. „Przemoc – zapobieganie i zwalczanie – Nowa Zelandia – od 1989 r. – publicystyka" (w tym wypadku chodzi o książkę *O winach i karach* Wiktora Osiatyńskiego). Ponadto niektóre teksty, np. *Talki w wielkim mieście* Moniki Piątkowskiej i Leszka Talki, to zbiory felietonów drukowanych uprzednio w prasie.

W tym miejscu należy zaznaczyć, że w prasie pojawiają się teksty niepublicystyczne, jak choćby wiersze czy powieści w odcinkach. Nie da się ich jednak automatycznie wyselekcjonować, w związku z tym pojedynczy tytuł prasowy jest przypisany jako całość do stylu. Oznacza to również, że niektóre tytuły publikujące w przeważającej mierze teksty literackie (np. *„Esensja"*) są w całości oznaczone jako literatura piękna, choć wiele pojedynczych artykułów tam ogłoszonych to raczej teksty publicystyczne.

Z kolei jeśli kanałem jest książka, pozostaje wybór między literaturą piękną, literaturą faktu, typem naukowo-dydaktycznym i informacyjno-poradnikowym. Na tym poziomie pierwszym krokiem było rozróżnienie między pierwszym z tych typów (literaturą piękną) a pozostałymi. Podstawowym kryterium było to, czy treścią utworu jest fikcja literacka. Jeśli tak, to tekst był klasyfikowany jako „literatura piękna". Ewentualne kolejne uszczegółowienia (proza, poezja, dramat) dodawano na podstawie przyjętych ogólnie w genologii kryteriów. Dodajmy tylko, że scenariusz filmu również klasyfikowano jako dramat.

Choć w teorii literatury definicja fikcji nie jest jednoznaczna[6], uznajemy ją za kryterium intuicyjnie łatwe do zastosowania. Jest ono zresztą stosowane powszechnie w innych korpusach narodowych, np. w BNC. Oczywiście zawsze pozostanie margines utworów z pogranicza reportażu i opowiadania.

Jeśli dzieło nie zawiera fikcji literackiej, to pozostaje dwojaki wybór: może należeć do szeroko pojętej literatury faktu lub mieć charakter dydaktyczny. Teksty o charakterze dydaktycznym przekazują w założeniu pewną wiedzę i (ewentualnie) umiejętności, dzielą się dalej na typ naukowo-dydaktyczny oraz informacyjno-poradnikowy. Istotnym kryterium rozróżniającym te dwa typy jest stopień hermetyczności informacji. Zakładamy, że tekst naukowo-dydaktyczny jest pisany przez specjalistę dla specjalisty, studenta lub ucznia, nosi językowe wykładniki obiektywizmu i przekazuje raczej fakty niż instruktaż. Pozostałe teksty o charakterze dydaktycznym zaliczamy do typu informacyjno-poradnikowego. Tu zaliczamy poradniki hobbystyczne (np. *Zwierzęta w moim domu* Hanny Jurczak-Gucwińskiej), poradniki dotyczące postępowania: zarówno poparte autorytetem psychologii czy medycyny (*Nastolatki i alkohol: strategie profilaktyczne w szkole* Mirosławy Wieczorek-Stachowicz), jak też oparte na ezoteryce, astrologii czy filozofii Wschodu (*Klucze do twojego wnętrza* Barbary Antonowicz-Wlazińskiej), książki o charakterze instruktażowym (*ABC tworzenia stron WWW* Bartosza Danowskiego) czy wreszcie przewodniki turystyczne, które również przekazują pewną wiedzę i stanowią instruktaż (co i w jaki sposób zwiedzić), i książki kucharskie. Dodajmy, że przekaz niektórych poradników nie musi być zgodny z ustaleniami nauki.

Jak widać, teksty informacyjno-poradnikowe to dość szeroka kategoria, częściowo definiowana negatywnie (tzn. jako teksty mające pewne cechy wspólne z tekstami naukowymi, ale niespełniające innych warunków, które pozwalają tekst zaliczyć do typu naukowo-dydaktycznego).

Z kolei teksty nieprzekazujące wiedzy lub umiejętności, ale niefikcjonalne zaliczaliśmy do typu oznaczanego jako literatura faktu. Wyznacznikiem tej klasy jest narracja niefikcjonalna. Literaturę faktu w naszym rozumieniu reprezentują zarówno książki reportażowe i podróżnicze (*Modlitwa o deszcz* Wojciecha Jagielskiego czy pisarstwo Arkadego Fiedlera), jak biografie i autobiografie (*Requiem dla ziemiaństwa* Mieczysława Jałowieckiego), dzienniki (np. Zofii Nałkowskiej) czy szerzej – literatura wspomnieniowa, a także esej historyczny (*Zniewolony umysł* Czesława Miłosza). Do tej kategorii zaliczyliśmy także pewne teksty, które dałoby się zakwalifikować jako książki popularnonaukowe z zakresu historii, traktując je jako rodzaj reportażu z przeszłości.

[6] „Nie jest rzeczą właściwą popadać w skrajności i twierdzić, że świat przedstawiony utworu literackiego pozwala się w całości zweryfikować przez zestawienie go ze znanym skądinąd wycinkiem rzeczywistości, lub odwrotnie, że świat ten w ogóle nie podlega żadnej tego rodzaju weryfikacji" (Głowiński i in. 1986: 58).

Wymienione typy nie wyczerpują wszystkich tekstów publikowanych w książkach. Dla tekstów niemieszczących się w zarysowanej typologii ze względu na to, że leżą na pograniczu kilku typów bądź są zbiorem tekstów przynależnych do różnych typów (*Żartem i pół serio* Henryka Markiewcza), albo dlatego, że przynależą do bardzo rzadkiej kategorii, więc nie warto dla niej tworzyć osobnego typu, jak np. zbiór kazań, czy wreszcie dlatego, że opisany schemat klasyfikacyjny zawodzi, utworzyliśmy odrębny typ „książka niebeletrystyczna nieklasyfikowana". Kategoria ta jest definiowana pozytywnie przez kanał (książka), negatywnie przez to, że nie jest to tekst fikcjonalny.

Kolejny typ tekstów, znów definiowany przede wszystkim negatywnie, to „inne teksty pisane". Generalnie są to teksty o charakterze już to perswazyjnym, już to informacyjnym, najczęściej krótkie. Kanałem tej kategorii tekstów są oprócz książek i gazet rękopisy, ulotki, druki ogłoszeniowe i reklamowe itp. Kategoria „inne pisane" ma cztery podkategorie: 1) typ urzędowo-kancelaryjny (najmocniej osadzony w genologii, grupujący przede wszystkim teksty normatywne), 2) teksty perswazyjne (ogłoszenia, reklamy, propaganda wyborcza i polityczna), 3) krótkie teksty instruktażowe oraz 4) listy[7].

Na koniec wreszcie mamy trzy różne typy tekstów mówionych. Przypomnijmy, że zaproponowana przez nas typologia – w przeciwieństwie do powszechnej praktyki w językoznawstwie korpusowym – nie zawiera kategorii „teksty mówione". Zamiast tego mamy trzy kategorie: mówione konwersacyjne, mówione medialne i quasi-mówione. Uznaliśmy bowiem, że różnice między nimi są na tyle istotne, że nie należy ich uznawać za jeden typ.

Zacznijmy od typu tekstów quasi-mówionych. Są to teksty wypowiedziane, ale utrwalone na piśmie (zapisane) przez kogoś innego niż zespół NKJP. Z zasady więc metodologia zapisu jest odmienna od przyjętej przez nas i co istotniejsze, teksty te zostały poddane wcześniej obróbce redakcyjnej. W znakomitej większości w naszym korpusie tę kategorię reprezentują protokoły obrad parlamentu. Oryginalne przemówienia są pozbawiane anakolutów, naturalnych dla tekstu mówionego błędów, przejęzyczeń itp. Jednak wydaje się, że zawierają wiele śladów spontanicznej mowy, choć oczywiście w wersji oficjalnej[8].

Teksty medialne i konwersacyjne są wiernie transkrybowane, a wszelkie ślady spontanicznej mowy są zachowywane. Te dwie grupy tekstów różnicuje natomiast sytuacja, w jakiej powstają. Teksty konwersacyjne są wypowiadane

[7] Wprawdzie wszystkie listy włączone do korpusu pochodzą na razie z wydań książkowych, jednak podobnie jak w wypadku internetowych wersji prasy kanał uznajemy tu za wtórny wobec typu.

[8] W powszechnym odczuciu protokoły parlamentu zawierają jedynie tzw. teksty wtórnie mówione (tzn. napisane po to, by następnie zostały odczytane publicznie). W rzeczywistości dotyczy to jedynie pewnych partii tekstu. Wiele fragmentów zawiera nawet elementy dialogu, zwracania się do konkretnych osób i odpowiadania im.

w sytuacji nieformalnej, całkowicie spontanicznie (por. rozdz. 4). Stopień spontaniczności w tekstach medialnych jest znacznie mniejszy, ze względu na oczywiste wymogi radia lub telewizji. Nie wiadomo też w poszczególnych przypadkach, do jakiego stopnia tekst jest wcześniej przygotowany.

Tak więc osią podziału tych tekstów jest najpierw metodologia zapisu (a w konsekwencji stopień, w jakim tekst odzwierciedla rzeczywisty tekst mówiony), dalej zaś opozycja sytuacji nieoficjalnej i oficjalnej, która decyduje o tym, czy mamy do czynienia z mową spontaniczną, czy też nie.

Na koniec przejdźmy do tekstów internetowych. Sieć WWW zawiera właściwie wszystkie typy tekstów z wyjątkiem mówionych[9], toteż sam fakt przekazu przez ten kanał nie może definiować żadnego typu. Jednak istnieją teksty, które mogą się pojawić jedynie w Internecie, a mianowicie blogi, czaty, listy dyskusyjne czy wreszcie strony WWW firmowe, urzędowe i prywatne. Jedynie w Internecie pojawia się Wikipedia (i inne strony powstałe na jej wzór). Zapewne dopiero wnikliwa analiza tekstów Narodowego Korpusu Języka Polskiego pozwoli stwierdzić, na ile są to teksty językowo odmienne od tego, co pojawia się w druku, ale już teraz można przedstawić kilka istotnych cech, które usprawiedliwiają wyróżnienie ich jako osobnego typu tekstów. Są one często pisane anonimowo, spontanicznie, przez osoby bez przygotowania w zakresie redakcji tekstu i w swoistej interakcji, wreszcie z zasady są to teksty krótkie bądź bardzo krótkie. Spontaniczność i interakcja (to znaczy odpowiedzi na zamieszczone uprzednio wypowiedzi) zbliżają blogi czy Usenet do tekstów mówionych. Ta ostatnia właściwość sprawiła, że teksty internetowe zostały podzielone na dwa typy: statyczne, jak firmowe strony WWW, i dynamiczne, jak blogi i czaty.

2.6. Przypadki graniczne

Niewątpliwie opisana powyżej procedura nie może za każdym razem dawać jednoznacznych wyników, choćby z tego powodu, że jak wspomnieliśmy wyżej, istnieją autorzy świadomie przekraczający granice stylów. W tym miejscu zasygnalizujmy tylko kilka wątpliwych przypadków.

W praktyce przyjęliśmy następującą procedurę kategoryzowania poszczególnych tekstów: obaj współautorzy niniejszego rozdziału niezależnie przypisywali książkę do danego typu. Ważną wskazówkę stanowiły dla nas hasła Biblioteki Narodowej, uznaliśmy bowiem kompetencję instytucji wyspecjalizowanej w klasyfikacji tematycznej książek. Tak więc, zasadniczo, jeśli np. hasło zawierało słowo „powieść", tekst był kategoryzowany jako literatura piękna, mutatis mutandis kategoryzowaliśmy teksty oznaczane jako „publicystyka". Następnie obie niezależnie

[9] Mamy tu na myśli transkrypcję – istnieje bowiem w Internecie bardzo wiele tekstów mówionych, ale jako pliki audio lub wideo.

przez nas zaproponowane kategoryzacje były porównywane. W przytłaczającej mierze były one zgodne. Niezgodności dotyczyły przypadków granicznych, mogły być też wynikiem błędu. W wypadku rozbieżności przyglądaliśmy się tekstowi uważniej, a następnie uzgadnialiśmy decyzję. Argumentacja towarzysząca podejmowaniu decyzji pozwalała uściślić kryteria.

2.7. Podsumowanie

Jakkolwiek przedstawiona typologia jest oparta na heterogenicznych kryteriach, odwołujących się raz do kanału, jakim tekst jest przekazywany, innym razem do społecznej konsytuacji tekstu, to znów do jego treści, to wydaje się, że poszczególne typy odpowiadają pewnej intuicji przeciętnego członka polskiej wspólnoty językowej. Choć powyższe kryteria mają charakter pozajęzykowy, to badania (Biber 1988, Górski i Łaziński 2010) wskazują na to, że w mniejszym czy większym stopniu odzwierciedlają one różnice wewnątrzjęzykowe. I tak fundamentalny jest podział na literaturę piękną fikcjonalną i resztę piśmiennictwa. Literatura naukowa, choć definiowana u nas jedynie przez charakter nadawcy i odbiorcy, stanowi (szczególnie w zakresie nauk eksperymentalnych) bardzo wyraźną i mocno zestandaryzowaną grupę tekstów. Stosunkowo bardziej wewnętrznie zróżnicowane są literatura faktu i typ informacyjno-poradnikowy.

2.8. Metadane

Każdy tekst jest opatrzony metadanymi zawartymi w nagłówku (por. rozdz. 10). W tym miejscu ograniczymy się do omówienia tych metadanych, które charakteryzują i klasyfikują tekst. Można więc powiedzieć, że mamy tyle komplementarnych typologii, ile informacji o tekście zawiera nagłówek. Dzięki ograniczaniu przeszukiwania do tekstów o metadanych spełniających pewien warunek (por. rozdz. 14) użytkownik może tworzyć podkorpusy na swój użytek. Równocześnie metadane pozwalają lokalizować źródło cytatu, czego wymaga zarówno rzetelność filologiczna, jak i prawo autorskie. Podstawowe dane to autor i tytuł tekstu. W wypadku czasopism i prac zbiorowych podawane są oba tytuły, tzn. tytuł całej publikacji i tytuł artykułu. To samo dotyczy wstępów i posłowi pisanych przez kogoś innego niż autor całego dzieła.

Wiele artykułów nie ma tytułu bądź ma tytuł konwencjonalny, powtarzający się wielokrotnie, jak np. „ogłoszenia drobne". W bardzo wielu wypadkach tekst jest anonimowy, jak to ma miejsce w wypadku tekstów internetowych, ale także znacznej liczby tekstów prasowych. Również wiele tekstów prasowych jest podpisywanych tzw. literałem, czyli rodzajem pseudonimu lub akronimu składającego się z kilku liter w nawiasie. W tym ostatnim wypadku oczywiście nie mamy do

czynienia z tekstem anonimowym, ale nie jest on – czego się zazwyczaj oczekuje – oznaczony imieniem i nazwiskiem autora.

Każdy tekst jest identyfikowany czasowo przynajmniej jedną z następujących dat: data wydania, data powstania, data pierwszego wydania, data pozyskania. Naturalnie w większości przypadków pojawia się jedynie pierwsza z dat. Gdy mamy do czynienia z kolejnym wydaniem książki, która powstała dużo wcześniej, metadane są uzupełniane o kolejne z wymienionych dat. Jeśli nie sposób ustalić daty powstania tekstu (co z zasady ma miejsce w wypadku druków ulotnych, a w niektórych wypadkach także tekstów internetowych), zamieszczamy datę pozyskania tekstu, którą należy traktować jako *terminus ante quem*. W wypadku książek podawana jest tylko data roczna, w wypadku artykułów prasowych – data dzienna bądź miesięczna.

Teksty prasowe mogą być opisywane numerem wydania, jeśli można było go ustalić. Książki i czasopisma zawierają w metadanych ISBN lub ISSN, a także miejsce wydania i nazwę wydawnictwa.

Oprócz wspomnianych wyżej danych – typu tekstu i jego kanału – nagłówek zawiera również odniesienie do Uniwersalnej Klasyfikacji Dziesiętnej (dotyczy to jedynie książek), oraz hasła bibliografii Biblioteki Narodowej. Te metadane charakteryzują tekst pod względem tematycznym.

Wreszcie niektóre z tekstów książkowych zawierają informację na temat płci autora. Wprawdzie jest to mniejszość korpusu, ale stanowi dobrą podstawę do badań nad cechami biolektów płciowych. Dobierając teksty do korpusu, zwracaliśmy uwagę na reprezentację obu płci, ale informacje o płci w metadanych mają na razie charakter eksperymentalny.

Nie wszystkie metadane są dostępne w wyszukiwarkach dla zewnętrznych użytkowników korpusu. Programy konkordancyjne NKJP wyświetlają znaczniki: tytułu, autora, daty, miejsca wydania oraz pierwodruku, wydawnictwa, typu tekstu oraz kanału. Wszystkie wymienione znaczniki pozwalają ograniczać przeszukiwanie, wydaje się jednak, że praktyczne znaczenie będą miały daty powstania bądź wydania tekstu, jego typ i kanał. Dane te, wspólne dla dużych ilości tekstów, pozwalają użytkownikowi tworzyć podkorpusy[10].

Pozostałe metadane nie są w tym zakresie użyteczne. Hasła bibliografii narodowej i Uniwersalnej Klasyfikacji Dziesiętnej przyjmują tak wiele wartości, że nie pozwalają w sensowny sposób ograniczać wyszukiwania.

[10] W wypadku książek wszystkie metadane były pozyskiwane ręcznie. W wypadku prasy ręcznie pozyskiwano tylko te metadane, które odnosiły się do całego tytułu prasowego, a więc jego tytuł, miejsce wydania, wydawcę, ISSN i hasła. Tytuł artykułu, autor, data i ewentualnie numer wydania były w większości wypadków pozyskiwane automatycznie, a następnie kontrolowane przez osoby dokonujące konwersji tekstów; pamiętajmy, że chodzi tu o dane dotyczące dziesiątków milionów tekstów. Z tego powodu ich wiarygodność jest nieco mniejsza niż w wypadku książek.

3

Reprezentatywność i zrównoważenie korpusu

Rafał L. Górski, Marek Łaziński

Jeśli ustalenia poczynione na podstawie danych pochodzących z korpusu mają być przenoszone na język jako całość, to korpus musi stanowić reprezentację „języka". Oczywiście nie da się stworzyć reprezentacji *langue* – bytu mentalnego. Jedyne, co można reprezentować, to *parole* – teksty przynależne do danego języka. Tak więc korpus to reprezentacja tekstów i do tekstów trzeba się odwoływać przy ustalaniu modelu reprezentatywności. Równocześnie jednak nie powinno się tracić z oczu faktu, że język to byt społeczny i społeczny aspekt jego funkcjonowania również musi być tutaj uwzględniony[1].

Jakkolwiek wszystkie korpusy referencyjne („narodowe") określane są przez swoich twórców jako zrównoważone, to ich budowa bywa oparta na bardzo odmiennych podstawach i w konsekwencji mocno zróżnicowana. Dość powiedzieć, że proporcje między tekstami książkowymi i prasowymi mogą się wahać od 4:1 (International Corpus of English) po 1:10 (FIDA PLUS – korpus słoweński). Podobnie stosunek tekstów fikcjonalnych do niefikcjonalnych może wynosić 1:28 (FIDA PLUS), 1:6,5 (International Corpus of English) albo 4:6 (czeski korpus SYN 2005).

Powyższe różnice wynikają oczywiście głównie z założeń przyjętych przy ustalaniu budowy korpusu, ale też są konsekwencją uwarunkowań zewnętrznych. Zauważmy, że FIDA, korpus języka niewielkiej społeczności językowej, jaką są Słoweńcy, jest bardzo słabo zrównoważony, po prostu dlatego, że w kraju liczącym 2 miliony mieszkańców nie powstaje tyle książek, by łatwo było stworzyć

[1] Świadomie mocno upraszczamy tutaj różnicę między Saussurowskim *langue*, a Chomskiańską kompetencją językową, między akcentem na społeczny i psychiczny wymiar języka, ponieważ te różnice nie są dla niniejszych rozważań istotne. Ważne w tym kontekście jest jedynie to, że językoznawstwo korpusowe dużo bardziej dostrzega społeczny niż psychiczny wymiar języka, por. Teubert 2005.

stumilionowy korpus, który nie wykazywałby dużej nierównowagi między tekstami ogłaszanymi w prasie i w książkach[2]. Nawet jednak w wypadku większych społeczności może być tak, że niektóre typy tekstów są niezbyt liczne, np. literatura naukowa z poszczególnych dziedzin. Przy ściśle przestrzeganych założeniach reprezentatywności stylistycznej i tematycznej mała liczba publikacji z jednej dziedziny mogłaby stanowić barierę dla wzrostu korpusu.

Na początek dokonajmy pewnego rozróżnienia. Zazwyczaj terminów reprezentatywność (ang. *representativeness*) i zrównoważenie (ang. *balance*) używa się zamiennie. My jednak rozumiemy te terminy odmiennie. Wyjaśnijmy je na początek czysto intuicyjnie. Reprezentatywność to odnoszenie się do jakiejś rzeczywistości istniejącej poza korpusem. Zrównoważenie zaś to dbałość o taką budowę korpusu, by żaden składnik na żadnym z poziomów nie dominował nad innymi[3]. Można sobie więc wyobrazić korpus reprezentatywny, ale nie zrównoważony. Można też sobie wyobrazić korpus zrównoważony, ale niereprezentatywny. Choć idealny korpus jest zarówno zrównoważony, jak i reprezentatywny, to jednak często zrównoważenie i reprezentatywność są postulatami, które się wzajemnie wykluczają.

Korpus może odzwierciedlać:

1. populację twórców tekstów,
2. populację tekstów,
3. produkcję tekstów,
4. recepcję tekstów,

przy czym nie jest to zapewne lista wyczerpująca. Ponadto każde z tych rozwiązań wymaga dopracowania szczegółów, tak więc każde z nich mogłoby być w inny sposób realizowane.

Rozważmy na początek kilka koncepcji stworzenia reprezentatywnego korpusu. Pierwszą możliwą zasadą tworzenia korpusu jest nie ustalać żadnych kryteriów reprezentatywności, ale skoncentrować się na zbudowaniu możliwie dużego i zróżnicowanego korpusu przypadkowo dobranych tekstów. Jeśli taki korpus będzie naprawdę duży i zróżnicowany, to należy się spodziewać, że – zgodnie z prawem wielkich liczb – będzie dobrze odzwierciedlać ogólną populację tekstów danego języka. W istocie takie niejawne założenie przyjmują ci wcale liczni językoznawcy, którzy posługują się wyszukiwarkami internetowymi jako korpusem. To rozwiązanie niesie ze sobą jednak niebezpieczeństwo tzw. samodobierającej się próby – ponieważ niektóre teksty da się znacznie łatwiej pozyskać niż inne, z góry można założyć, że taka próba będzie niereprezentatywna.

[2] Z tej samej przyczyny korpus słoweński jako jeden z niewielu narodowych dopuszcza teksty tłumaczone na równi z oryginalnymi słoweńskimi.

[3] Bardzo ogólnikową, choć trafną definicję terminu „balance" przedstawia Atkins i in. (1992): „By 'balanced corpus' is meant (apparently) a corpus so finely tuned that it offers a manageably small scale model of the linguistic material which the corpus builders wish to study".

Drugie rozwiązanie to ustalenie pewnej liczby typów tekstów, a następnie wypełnienie tych typów równą liczbą słów. Takie rozwiązanie zostało przyjęte w pierwszym polskim korpusie elektronicznym budowanym na potrzeby *Słownika frekwencyjnego współczesnej polszczyzny* (por. Kurcz i in. 1990). Korpus tego rodzaju można nazwać idealnie zrównoważonym, ale na pewno nie reprezentatywnym. Zauważmy bowiem, że ten sam udział mają w nim teksty prasowe i naukowe. Tymczasem zasięg i społeczne oddziaływanie obu tych typów tekstów są nieporównywalne. To tak, jakby przeprowadzić badanie opinii publicznej na podstawie próby z identyczną liczbą osób o dochodach niskich, średnich oraz milionerów. Rozwiązaniem, które narzuca się jako najoczywistsze, jest reprezentatywność rozumiana jako reprezentacja ogólnej populacji tekstów (liczonych w tytułach). Jest to założenie dobrze umotywowane teoretycznie i łatwo uchwytne intuicyjnie – korpus reprezentuje ogół tekstów drukowanych[4]. Również ustalenie populacji jest łatwe: zasadniczo wszystkie druki są skatalogowane przez Bibliotekę Narodową. Jednak i to założenie jest obciążone pewną zasadniczą wadą. Otóż przy ścisłym zastosowaniu otrzymamy korpus, który będzie co prawda reprezentatywny, ale zapewne nie zrównoważony. Z pewnością będzie wykazywał bardzo duży udział publikacji naukowych, tekstów o przecież dość specyficznym języku (nie tylko pod względem słownictwa); zapewne też korpus taki będzie wykazywał wyraźną przewagę tekstów prasowych. Istotniejsze jest jednak co innego: tworząc reprezentację tekstów ignorujemy całkowicie społeczny i psychiczny charakter języka. Przy takim podejściu bowiem interesują nas jedynie teksty i nie dostrzegamy tych, którzy je tworzą i dekodują, czyli komunikujących się ludzi. Język jawi się więc jako suma tekstów zupełnie oderwanych od użytkowników języka.

Reprezentatywność rozumiana jako reprezentacja twórców tekstów jest oczywista dla podkorpusu tekstów mówionych. Twórcą tekstu mówionego jest każdy członek danej wspólnoty językowej, próbkowania można by dokonać dokładnie w ten sam sposób w jaki losuje się respondentów do badań socjologicznych. Zastosowanie tego rozwiązania do teksów drukowanych natrafia na zasadniczą trudność: większość ludzi nie publikuje żadnych tekstów. Tworzenie poprawnie zbudowanej reprezentacji osób ogłaszających swoje teksty drukiem jest przedsięwzięciem ryzykownym. Ponadto – co istotniejsze – osoby takie stanowią znikomą część społeczeństwa. W tym miejscu jednak trzeba powiedzieć, że sytuacja zmienia się niezwykle szybko. Otóż wraz z pojawieniem się Web 2.0 coraz więcej osób ogłasza swoje teksty w Internecie. Zatem to rozwiązanie – jeszcze niedawno traktowane jako całkowicie hipotetyczne – powoli staje się godne rozważenia.

Kolejny model reprezentatywności to reprezentacja percepcji języka przez daną wspólnotę językową. W praktyce polegałoby to na tym, że staramy się

[4] Metody próbkowania tekstów na potrzeby tak rozumianej reprezentatywności opisuje Biber 1993.

w korpusie odzwierciedlić strukturę czytelnictwa w Polsce. Innymi słowy, im bardziej czytany jest dany typ tekstów, tym większy ma on udział w korpusie. Jakkolwiek nie otrzymujemy w efekcie zrównoważonego korpusu, zaletą rozwiązania jest to, że nie tracimy z pola widzenia użytkownika języka.

Nim jednak zajmiemy się odwzorowaniem w korpusie percepcji języka, musimy z góry zgodzić się na to, że nie da się ze względów czysto technicznych utrzymać proporcji między tekstami, które docierają do użytkownika języka przez mowę i pismo. Pozyskanie tekstów mówionych jest znacznie bardziej skomplikowane i kosztowne. W związku z tym przyjmujemy całkowicie arbitralnie, że jedynie 10% tekstów korpusu to teksty mówione, z czego większość to teksty quasi-mówione. Ta wielkość wynika jedynie z oceny możliwości pozyskania tekstów mówionych. Tak więc 10% to mówione i quasi-mówione[5], a 90% korpusu to teksty pisane i to ich recepcję chcemy badać, by ustalić kryteria reprezentatywności.

Mimo swoich zalet podejście percepcji tekstów stwarza wiele problemów metodologicznych i praktycznych. Powiedzmy na początku, że nie staramy się odzwierciedlić w korpusie recepcji języka przez mitycznego „przeciętnego czytelnika", ale raczej przez całe społeczeństwo.

W ideale powinniśmy przeprowadzić badanie socjologiczne, które odpowie nam na pytanie, ile słów przeciętny czytelnik „konsumuje" w poszczególnych typach i kanałach tekstów, a następnie proporcjonalnie do tego przyznać tym typom i kanałom odpowiedni udział w korpusie[6].

Recepcja tekstów (czytelnictwo) nie zawsze jest wprost proporcjonalna do ogólnej populacji tekstów i ich produkcji. O ile wspomniana wyżej reprezentatywność ogólnej populacji tekstów dotyczy tytułów (daje każdemu tytułowi taką samą moc statystyczną), to badanie recepcji dotyczy obiegu tekstów, np. w badaniach preferencji czytelniczych. Oczywiście w takim rozumieniu reprezentatywności jeden tytuł publikacji specjalistycznej jest dla struktury reprezentatywności mniej istotny niż tytuł gazety, a nawet jej pojedynczy numer.

Powtórzymy: chcemy, żeby komponent pisany podkorpusu zrównoważonego NKJP odzwierciedlał recepcję polszczyzny pisanej. W tym celu staramy się zrekonstruować strukturę czytelnictwa w Polsce.

Dla książek najważniejsze źródła to cykliczne raporty Instytutu Książki i Czytelnictwa Biblioteki Narodowej: Straus i Wolff 1996a, b, 1998, 2000, 2002, 2004, 2006, Straus i in. 2008. Nie dają one wprost odpowiedzi na nasze pytania, badają bowiem raczej zachowania kulturowe Polaków. Podstawą dla nich są ankiety, w których padają pytania o czytane bądź nieczytane typy tekstów oraz ogólnie o liczbę książek przeczytanych w ciągu roku. Typy tekstów, które respondenci

[5] Definicja tekstów quasi-mówionych jest podana w rozdz. 2.

[6] O ile nam wiadomo, spośród lingwistów korpusowych tylko twórcy Czeskiego Korpusu Narodowego zastosowali konsekwentnie tę procedurę.

mają wskazać, nie do końca pokrywają się z naszą typologią (por. rozdz. 2). Zaletą tych raportów jest natomiast duża wiarygodność. Badania są bowiem prowadzone przez wyspecjalizowany instytut badania opinii, a same wyniki są dość stabilne.

Kolejne źródła statystyczne dotyczą czasopism. Są to dane Polskiego Badania Czytelnictwa[7] oraz Związku Kontroli Dystrybucji Prasy[8]. Pierwsze z tych badań daje informację o tym – w pewnym uproszczeniu – ile osób zna dane czasopismo. Drugie badanie określa wielkość sprzedanego nakładu czasopisma. Nie są to dokładnie te dane, których potrzebujemy, niemniej zawierają one (choćby pośrednio) bardzo istotną informację na temat preferencji czytelniczych Polaków.

Od korpusu reprezentatywnego należy oczekiwać odzwierciedlenia języka w całej jego różnorodności, tak by każdy typ tekstu miał jakiś, choćby niewielki udział w korpusie. Stąd wzięła się w naszej typologii heterogeniczna kategoria „inne teksty pisane" (por. p. 2.2). Nie sposób określić recepcji tekstów, które się na nią składają, wydaje się zresztą, że w wypadku ustaw, czy orzecznictwa sądowego jest ona minimalna. Przyjmujemy arbitralnie, że na 90% wszystkich tekstów pisanych udział tekstów niesklasyfikowanych wyniesie 3%. Kolejna arbitralna decyzja to 7% udziału tekstów internetowych. Należy w tym miejscu zastrzec, że od momentu rozpoczęcia prac nad Narodowym Korpusem Języka Polskiego (2007) i nad innymi korpusami referencyjnymi w Polsce (koniec lat dziewięćdziesiątych) zasięg Internetu znacząco się zwiększył. Jednak nie próbowaliśmy ustalić proporcji tekstów internetowych i „papierowych" ze względu na to, że trudno posłużyć się tu tą samą metodologią co w wypadku tekstów drukowanych. Ponadto – jak wspomnieliśmy w rozdz. 2 – jako teksty internetowe traktujemy tylko część spośród tekstów pojawiających się w tym medium, internetowe wydania prasy klasyfikujemy jako kanał prasowy. Jak łatwo zauważyć, te dwie arbitralnie przyjęte wielkości obejmują kolejne 10%, które dodajemy do wcześniej wydzielonych 10% tekstów mówionych i quasi-mówionych.

Pozostałe 80% korpusu to teksty, dla których prymarnym medium jest prasa lub książka. Mimo że staramy się zapewnić reprezentację typów tekstów, a nie ich kanałów, uznajemy, że prymarnie to, co jest ogłaszane w prasie, stanowi „publicystykę i krótkie wiadomości prasowe", natomiast książki stanowią prymarny kanał dla pozostałych typów (literatury pięknej fikcjonalnej, literatury faktu, tekstów naukowo-dydaktycznych i informacyjno-poradnikowych). Ustalenie proporcji między recepcją prasy i książek pozwala ustalić udział procentowy publicystyki i krótkich wiadomości prasowych w korpusie.

Wiemy dość precyzyjnie, ile książek czyta przeciętny Polak, co pozwala w przybliżeniu powiedzieć, jaka jest średnia roczna objętość tekstów przezeń czytanych. Jeśli chodzi o gazety, to mamy dane jeszcze precyzyjniejsze – ile gazet

[7] http://www.pbczyt.pl/.

[8] http://www.zkdp.pl/.

jest kupowanych. Niestety trudno dotrzeć do informacji na temat tego, jaka część zakupionego egzemplarza gazety realnie jest czytana. Można bowiem przyjąć, że książka jest czytana od deski do deski (co nie musi być prawdą w wypadku poradników, a na pewno nie w wypadku encyklopedii), jednak w gazecie czyta się jedynie wybrane artykuły. W zasadzie jedyna informacja, na jaką udało się nam trafić, to artykuł o czytelnictwie prasy: Makarenko 2001[9]. Wiemy z niego, że statystyczny Polak spędza na lekturze prasy 20 minut dziennie. Przyjmując więc, że czyta się z prędkością 250 słów na minutę (Kurcz i Polkowska 1990: 13) otrzymujemy 5000 słów dziennie, czyli 1 825 000 słów rocznie przeczytanych w prasie przez jednego Polaka. Wiedząc z badań Biblioteki Narodowej, że Polak[10] czyta średnio 8 książek rocznie i przyjmując, że przeciętna książka liczy 70 000 słów (dane z korpusu) otrzymujemy 560 000 słów rocznie przeczytanych w książkach, więc 3,2 razy mniej niż w prasie.

Jeśli roboczo nazwiemy „książkowymi" te typy tekstów, dla których prymarnym kanałem jest książka, to nasz korpus miałby w przybliżeniu strukturę opisaną w tab. 3.1.

Tabela 3.1. Hipotetyczna struktura korpusu reprezentatywnego

Typ tekstu	Udział
Publicystyka i krótkie informacje prasowe	56%
Teksty książkowe	24%
Inne teksty pisane	10%
Mówione	10%

W tym miejscu jednak przypomnijmy o drugim z postulatów: korpus powinien być nie tylko reprezentatywny, ale i zrównoważony. Uznajmy tedy, że zrównoważenie oznacza tyle, iż żaden z typów tekstów nie obejmuje więcej niż połowę korpusu. W związku z tym zmniejszamy udział publicystyki do 50%, przy czym 49% ma pochodzić z prasy a 1% z książek publicystycznych.

Wprawdzie wydawać by się mogło, że jest to decyzja czysto arbitralna, ale nie chcemy w korpusie dawać aż takiej przewagi prasie nad książkami również ze względu na to, że książki na pewno częściej zostają w pamięci, ich lektura bywa też zadawana i sprawdzana w szkole. Uwzględniając większą istotność tekstów książek dla świadomości językowej i kulturowej, proponujemy, by publicystyka

[9] Ten artykuł wskazał nam p. dr Ryszard Filas z Ośrodka Badań Prasoznawczych UJ, za co Mu w tym miejscu dziękujemy.

[10] Bierzemy tu pod uwagę tylko respondentów z wyższym i niepełnym wyższym wykształceniem. Zwykle blisko połowa respondentów przyznaje, że nie przeczytała żadnej książki w ciągu roku. Gdybyśmy mieli wziąć pod uwagę całą populację, to liczba przeczytanych książek byłaby znikoma.

i krótkie wiadomości prasowe miały w korpusie udział tylko dwukrotnie większy niż typy książkowe. W wyniku tego ograniczenia przybliżona budowa korpusu wygląda jak w tab. 3.2.

Tabela 3.2. Hipotetyczna struktura korpusu zrównoważonego

Typ tekstu	Udział
Publicystyka i krótkie wiadomości prasowe	50%
Typy książkowe	30%
Inne (pisane nieklasyfikowane, urzędowe, Internet)	10%
Mówione	10%

Jak wspomnieliśmy, istnieją dwa wiarygodne źródła dotyczące czytelnictwa prasy: Polskie Badanie Czytelnictwa oraz raporty Związku Kontroli Dystrybucji Prasy. Obie instytucje badają większość tytułów, ale nie wszystkie. Pierwsze źródło nie informuje, czy respondent rzeczywiście czyta dany tytuł, czy tylko go zna. Drugie źródło informuje o sprzedaży prasy, a można założyć z niewielkim ryzykiem błędu, że osoba, która kupuje gazetę – czyta ją, i przyjąć, że liczba czytelników jest równa liczbie sprzedanych egzemplarzy.

To założenie pozwoli nam ustalić proporcje między różnymi typami tekstów w obrębie prasy. W tym miejscu zaznaczmy, że pierwotnie staraliśmy się zaadaptować typologie wypracowane przez prasoznawców, okazały się jednak nieprzydatne. Po pierwsze dlatego, że jeden tytuł zazwyczaj przynależy do wielu kategorii. Po drugie, typologie te mają charakter w dużej mierze tematyczny, tymczasem my – o czym niżej – staramy się raczej o zrównoważenie tematyczne niż reprezentatywność w tym zakresie.

Wydaje się natomiast, że tym, co naprawdę różnicuje prasę, są proporcje między krótkimi wiadomościami i publicystyką. Te pierwsze są zasadniczo spotykane jedynie w dziennikach, częściowo też w prasie lokalnej. Publicystyka zajmuje zarówno łamy gazet[11] (mniejszą część), jak też pozostałych czasopism. Ustalenie proporcji między dziennikami a pozostałymi periodykami pozwala choćby w przybliżeniu ustalić proporcję między tymi dwoma składnikami typu określanego przez nas ogólnie jako „publicystyka i krótkie wiadomości prasowe".

Aby ustalić te proporcje, weźmy 180 tytułów o najwyższym sprzedanym nakładzie. Przemnóżmy średni nakład przez liczbę wydań w roku. Po zsumowaniu okaże się, że liczby egzemplarzy dzienników i czasopism sprzedanych w ciągu roku są zbliżone i wynoszą odpowiednio 1,3 i 1,1 miliarda. Pozwala to ustalić proporcje między tymi rodzajami prasy na odpowiednio 54% i 46%. Zwracamy jednak uwagę, że wszystkie teksty prasowe są oznaczone w korpusie jako jeden typ.

[11] Za gazetę uznajemy wydawnictwo ciągłe ukazujące się częściej niż raz w tygodniu.

Z kolei udział poszczególnych typów tekstów określanych tutaj roboczo jako „książkowe" ustalamy na podstawie raportów Biblioteki Narodowej. Jeśli znormalizujemy te dane, otrzymamy procentowy udział poszczególnych typów tekstów. Podkreślmy wszakże, że udział ten nie jest proporcjonalny do objętości czytanego tekstu, ale do liczby osób deklarujących czytanie książek przynależnych do danego typu. Innymi słowy, im więcej osób deklaruje czytanie którejś z kategorii książek, tym większy jej udział w korpusie. Można na to popatrzeć z drugiej strony: mniej ludzi deklarujących czytanie danego typu tekstu oznacza jego mniejszy zasięg społeczny, a w konsekwencji mniejszy udział w korpusie. W ten sposób teksty w społecznym odbiorze marginalne są słabiej reprezentowane w korpusie niż teksty o dużym zasięgu społecznym.

Zaznaczmy, że choć w niniejszej klasyfikacji przyjmujemy płaską strukturę typów tekstów, to jednak podstawowym rozróżnieniem jest podział na teksty fikcjonalne i niefikcjonalne. Podział ten jest ugruntowany w tradycji filologicznej (choćby przez to, że te drugie rzadko są przedmiotem zainteresowania literaturoznawców), w handlu księgarskim i ma głębokie uzasadnienie ściśle językowe, czego dowodzi praca Górski i Łaziński 2010. Upoważnia nas to do ustalenia budowy komponentu książkowego w dwu krokach: najpierw podziału wzdłuż osi teksty fikcjonalne i niefikcjonalne, a potem głębszego podziału komponentu tekstów niefikcjonalnych.

Taka, a nie inna procedura wynika również z sygnalizowanego wcześniej braku jednoznacznej odpowiedniości typów książek wyróżnianych w cytowanych powyżej raportach BN i przyjętej przez nas typologii. O ile kategorie pojawiające się w ankietach dają się z jednym wyjątkiem łatwo zakwalifikować do fikcji bądź tekstów niefikcjonalnych, o tyle przy ustalaniu budowy podkomponentu niefikcjonalnego trzeba dokonywać pewnych przybliżeń. Czytanie różnych gatunków literatury pięknej deklaruje ponad połowa (52,5%), a czytanie książek non fiction niemal połowa respondentów (47,3%). Pozwala to przyjąć niewielką przewagę tej pierwszej.

Jeśli chodzi o książki non fiction, obraz nie jest tak jednoznaczny. Problem może stanowić na przykład wymieniana w ankietach kategoria „ezoteryka i ufologia", jest ona jednak marginalna (czytanie tego typu książek deklaruje tylko 1% ankietowanych) lub kategoria „religijne" (mniej marginalna). Obie wymienione kategorie nie znajdują odpowiedników w naszej typologii. Jeśli w danych z ankiet BN pominiemy kategorie wyróżniane na podstawie kryteriów czysto tematycznych, a nie stylistycznych, możemy na podstawie tych ankiet zaproponować podział książek non fiction jak w tab. 3.3.

Po uwzględnieniu wszystkich dotychczasowych parametrów otrzymujemy wzorcową budowę korpusu przedstawioną w tab. 3.4 i na rys. 3.1.

Tabela 3.3. Proponowany podział książek non fiction

Typ tekstu	Udział wśród tekstów non fiction
Informacyjno-poradnikowe	40%
Literatura faktu	40%
Naukowo-dydaktyczne[a]	10%
Publicystyka[b]	10%

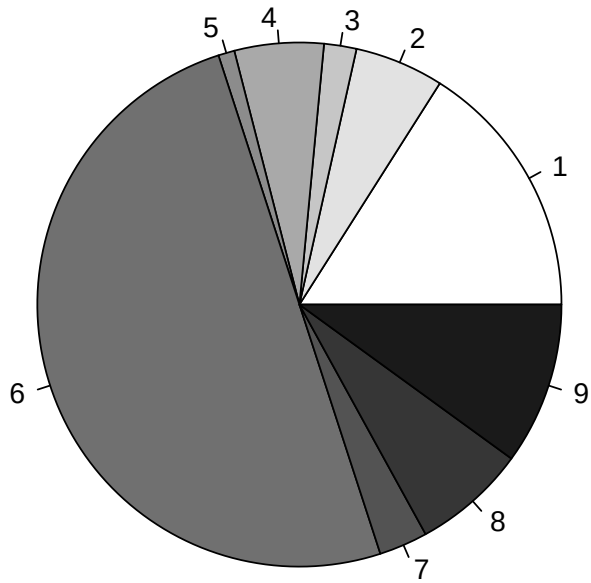
[a] W ankietach BN pada pytanie o lekturę książek „fachowych". Kategoria ta ma charakter raczej tematyczny i funkcjonalnie pokrywa się w znacznej mierze ze stosowanym przez nas typem naukowo-dydaktycznym.
[b] Raporty BN łączą książkę publicystyczną i eseistykę w jedną kategorię.

Tabela 3.4. Wzorcowa budowa korpusu

Typ tekstu	Udział procentowy w korpusie
Publicystyka i krótkie wiadomości prasowe	50,0%
Literatura piękna	16,0%
Literatura faktu	5,5%
Typ informacyjno-poradnikowy	5,5%
Typ naukowo-dydaktyczny	2,0%
Inne teksty pisane	3,0%
Książka niebeletrystyczna nieklasyfikowana	1,0%
Teksty konwersacyjne, mówione medialne i quasi-mówione razem	10,0%
Teksty internetowe statyczne i dynamiczne razem	7,0%

Powróćmy do sygnalizowanej wyżej hierarchii typów tekstów. Otóż przedstawioną powyżej budowę mają korpus 1M (jednomilionowy, ręcznie znakowany; por. tab. 5.1 na s. 53) oraz demo 100M (wersja demo korpusu zrównoważonego liczącego 100 milionów słów). W korpusie zrównoważonym liczącym 300 milionów segmentów (a więc ok. 250 milionów tradycyjnie rozumianych słów; por. p. 6.2.2) została zachowana proporcja głównych typów tekstów: prasa, beletrystyka, książki non fiction, inne teksty pisane, wszystkie trzy rodzaje tekstów mówionych. Nie zostało natomiast zachowane zróżnicowanie wewnątrz kategorii „książka non fiction". Okazało się bowiem, że w naszym korpusie nieco za słabo jest reprezentowany typ poradnikowo-informacyjny. Stosunkowo niewielki niedobór tego rodzaju tekstów spowodowałby konieczność znacznego zmniejszenia

Rysunek 3.1. Wzorcowa budowa korpusu



1. literatura piękna
2. literatura faktu
3. typ naukowo-dydaktyczny
4. typ informacyjno-poradnikowy
5. książka niebeletrystyczna niesklasyfikowana
6. publicystyka i krótkie wiadomości prasowe
7. inne teksty pisane
8. teksty internetowe statyczne i dynamiczne
9. teksty konwersacyjne, mówione medialne oraz quasi-mówione

całego korpusu zrównoważonego. Uznaliśmy, że ścisłe zachowanie tych proporcji ma dla użytkownika korpusu znacznie mniejsze znaczenie niż jego wielkość. Dlatego powiększyliśmy nieco udział książek reprezentujących typ naukowo-dydaktyczny, jako funkcjonalnie najbliższy typowi informacyjno-poradnikowemu, kosztem tego ostatniego.

Należy jednak podkreślić, że korpus 300M nie ma innej budowy niż 1M oraz 100M. Mniejsze korpusy są staranniej zrównoważone jedynie w tym sensie, że zadbano nie tylko o zachowanie proporcji książek non fiction względem reszty korpusu, ale także o proporcje poszczególnych typów i gatunków tych książek.

Mówiąc o zrównoważeniu, powinniśmy wziąć pod uwagę nie tylko typy tekstów, ale także zrównoważenie tematyczne. Zwracamy uwagę na to, że jeśli

chodzi o tematykę tekstów w korpusie, dbamy jedynie o zróżnicowanie i zrównoważenie, nie staramy się w żaden sposób zapewnić reprezentatywności korpusu pod względem tematyki. Innymi słowy, gdy dokonujemy selekcji tekstów, najpierw włączamy tekst, którego tematyka albo nie jest dotąd reprezentowana, albo jest reprezentowana słabiej niż pozostałe.

Dla uzyskania lepszego zrównoważenia niekiedy w korpusach stosuje się próbkowanie pojedynczych tekstów. O ile ten zabieg wydaje się celowy w niewielkich korpusach, gdzie różnice objętości poszczególnych książek mogą być istotne, o tyle w dużym korpusie największa nawet książka stanowi jego znikomą część. Fragmenty książek są włączane do korpusu jedynie wtedy, gdy wydawca zgodził się na przekazanie pojedynczych rozdziałów. Książki nadmiarowe pod względem stylu czy tematyki były odrzucane w całości. Natomiast w wypadku czasopism konieczne było próbkowanie po to, by zmieścić w korpusie zrównoważonym reprezentację każdego tytułu.

Zrównoważenie stylistyczne jest nieodzownym warunkiem uznania zbioru tekstów za referencyjny korpus narodowy, tzn. odzwierciedlający świadomość językową przeciętnego użytkownika danego języka. Drugim ważnym parametrem struktury korpusów narodowych jest chronologia tekstów. Jeśli spojrzymy na ramy czasowe różnych korpusów narodowych, okaże się, że różnorodność jest tu podobna jak w wypadku zróżnicowania stylistycznego. Wspomniany na początku korpus SYN2005, tworzony w ramach Czeskiego Korpusu Narodowego, zawiera teksty wyłącznie z XXI w. Korpus SYN2000 gromadzi prasę od roku 1990 oraz książki autorów urodzonych po roku 1880. British National Corpus (powstały w początkach lat dziewięćdziesiątych ubiegłego wieku) robi dla powieści wyjątek w chronologii (która zakłada, że teksty nie powinny być wcześniejsze niż z roku 1975) i uwzględnia powieści pisane od roku 1964. W narodowym korpusie słoweńskim FIDA PLUS teksty z obecnego tysiąclecia stanowią 70%, a niemal wszystkie pozostałe teksty powstały po roku 1990. Jedynie 0,04% tekstów pochodzi z lat 1970–1989. Dla kontrastu Narodowy Korpus Języka Rosyjskiego ustala granicę czasową na początek XIX wieku, włączając twórczość Nikołaja Karamzina. Charakterystyczne, że rozciągłość czasowa korpusu referencyjnego jest zwykle wprost proporcjonalna do udziału w nim literatury pięknej – w korpusie słoweńskim mamy 3,5% beletrystyki, a w korpusie rosyjskim aż 40%.

Które lata powinny być granicami chronologicznymi dla zrównoważonego i reprezentatywnego korpusu polszczyzny? Należałoby zapewne znaleźć złoty środek między nowoczesnością korpusów zachodnich (w tym zachodniosłowiańskich) oraz silną tradycją literaturocentyczną leksykografii i korpusów rosyjskich (notabene leksykografia polska co najmniej do lat osiemdziesiątych XX wieku nawiązywała do podobnej tradycji silnego autorytetu kulturowego klasyki literatury

pięknej). Korpus nie powinien obejmować tylko tekstów powstałych po roku 1990, ponieważ w takim razie ograniczylibyśmy istotnie tradycję literacką nawet na poziomie podstawowej listy lektur powieściowych szkoły średniej. Mniej oczywista jest reprezentacja tekstów prasowych sprzed kilkudziesięciu lat ze względu na ogromne zmiany technologiczne, społeczne i polityczne, których umowną granicą jest rok 1989. Uznaliśmy jednak, że ograniczenie się do 20 ostatnich lat uniemożliwiłoby w korpusie badania leksykalne z uwzględnieniem rozwoju i zmian znaczeń.

Dla korpusu zrównoważonego przyjęliśmy za cezurę rok 1945, według Zenona Klemensiewicza (1985) wyznaczający granice polszczyzny współczesnej. Jedynie dla tekstów literackich przesuwamy te granicę na początek XX wieku, bo przecież dla przeciętnego Polaka powieści Dołęgi-Mostowicza czy Choromańskiego, a nawet Żeromskiego, są tekstami bliskimi polszczyźnie współczesnej (naturalnie w wydaniach z ortografią po reformie z 1936 roku). Oczywiście wiek XX nie jest w korpusie reprezentowany równomiernie. Najwięcej tekstów (80%) powstało po roku 1990, 15% w latach 1945–1990, a tylko 5% przed rokiem 1945.

Wszystkie opisane dotychczas zasady reprezentatywności tekstów dotyczą w Narodowym Korpusie Języka Polskiego podkorpusu zrównoważonego liczącego 300 milionów segmentów. Według tych samych kryteriów wybrano teksty do korpusu liczącego milion słów, który został ręcznie oznakowany i ujednoznaczniony, a następnie wyposażony w znaczniki syntaktyczne i semantyczne na różnych poziomach (por. rozdziały 6–9). Trzeba pamiętać, że 300 milionów segmentów w tekstach o zrównoważonej i zróżnicowanej tematycznie strukturze stanowi zaledwie kilkanaście procent całej zawartości NKJP, która gromadzona była przez ponad 2 lata z myślą o zrównoważeniu dużego podkorpusu, ale bez odrzucania tekstów prasowych czy książkowych nadmiarowych w danym typie stylistycznym czy chronologicznym. Najstarsze teksty w tym wielkim korpusie pochodzą z przełomu XVIII i XIX wieku, czyli z czasów, które powszechnie uznaje się za początek doby nowopolskiej w historii polszczyzny, a mówiąc bardziej obrazowo – za początek literatury, którą przeciętny polski maturzysta jest w stanie czytać bez specjalistycznego słownika. Korpus zrównoważony będzie zapewne podstawowym narzędziem do badań leksykograficznych, także do wyciągania wniosków socjolingwistycznych czy nawet socjologicznych z kolokacji. Takie wnioski pochopnie byłoby wyciągać z całości niezrównoważonego stylistycznie materiału liczącego ponad 1,5 miliarda słów otagowanych i ujednoznacznionych gramatycznie, który to materiał znakomicie nadaje się do wyszukiwania skomplikowanych kategorii gramatycznych, do kwerend leksykalnych słów najrzadszych.

4

Język mówiony w NKJP

Piotr Pęzik

4.1. Język mówiony a reprezentatywność korpusu

Uwzględnienie języka mówionego w Narodowym Korpusie Języka Polskiego nie wymaga długiego uzasadnienia. Korpus narodowy ma charakter referencyjny i jako taki powinien stanowić źródło informacji o tak ważnej odmianie języka, jaką jest ogół języka mówionego, tym bardziej, że to właśnie język mówiony stanowi największą pod względem ilości naturalnie występujących wypowiedzi odmianę języka. Wydaje się, iż mimo ostatniego wzrostu poziomu piśmiennictwa spowodowanego upowszechnieniem się Internetu jako platformy komunikacji językowej, rodzimi użytkownicy polszczyzny dużo więcej mówią, niż piszą, oraz dużo częściej są odbiorcami języka mówionego niż czytelnikami tekstów[1]. Język mówiony zasługuje więc na szczególne uwzględnienie w korpusie referencyjnym, bez względu na to, czy jego głównym kryterium doboru tekstów jest poziom odbioru, czy też produkcji danej odmiany języka.

Szczególną rolę komponentu danych mówionych w korpusie referencyjnym uzasadniają także wyjątkowe własności w warstwie leksykalnej, frazeologicznej, składniowej oraz pragmatycznej języka mówionego. Kilka przykładów wyjątkowości języka mówionego można znaleźć w dalszej części tego rozdziału, ale warto tu wspomnieć o tym, jak ilościowa analiza języka może zależeć od rodzaju danych, na podstawie których jest przeprowadzana. Jedną z podstawowych technik ilościowej analizy korpusu są listy frekwencyjne wyrazów. Czy na podstawie listy frekwencyjnej wyrazów wygenerowanej z całości danych NKJP można wskazać

[1] Twierdzenie, że więcej języka słyszymy, niż czytamy, wydaje się szczególnie prawdziwe, jeżeli przez pojęcie języka mówionego rozumiemy kanał, a nie typ języka. Trudno bez wiarygodnych danych stwierdzić, czy odbieramy więcej języka mówionego w jego mniej lub bardziej swobodnych odmianach niż czytanego.

najczęstszy wyraz w polszczyźnie? Niemal każdy językoznawca korpusowy zajmujący się angielszczyzną wie, że najczęstszym wyrazem w ogromnej większości korpusów języka angielskiego jest przedimek określony *the*[2]. Zgodnie z prawem Zipfa o rozkładzie frekwencji słów, częstość *the* jest zazwyczaj (co najmniej) dwukrotnie większa od częstości kolejnego słowa na liście. Niestety dużo trudniej wskazać taki wyraz w języku polskim. W języku pisanym jest to zazwyczaj przyimek *w*, ale w podkorpusie danych mówionych konwersacyjnych w NKJP ten sam wyraz plasuje się na piątym miejscu listy frekwencyjnej (tab. 4.1). Ze względu na swoją wieloznaczność oraz funkcje dyskursywne wyrazy *to*, *nie* oraz *no* są dużo częstsze w języku mówionym niż w tekstach pisanych. Jeżeli przyjąć, że więcej mówimy niż piszemy, może to również oznaczać, że przyimek *w* nie jest najczęstszym polskim wyrazem. Przykład ten potwierdza, iż wyników analizy ilościowej przeprowadzanej na podstawie korpusu zawierającego tylko próbki języka pisanego nie można rozciągać na język mówiony, nawet na tak prostym poziomie, jakim są frekwencje pojedynczych leksemów. Dlatego też w NKJP znalazły się próbki języka mówionego zebrane przez grupę PELCRA (z lat 2000–2005) oraz zebrane specjalnie na potrzeby NKJP dane mówione oraz konwersacyjne z lat 2008–2010. Aktualność tych danych widać w wynikach zapytań o leksemy, które w ciągu kilku ostatnich lat nabrały nowego znaczenia w języku mówionym. Przykładem tego mogą być zarejestrowane w NKJP użycia wyrazów *ogarnąć/ogarniać się* oraz *ogarnięty*, które są używane zwłaszcza przez ludzi młodych odpowiednio w znaczeniach *opanować/poradzić sobie z czymś*, *zrozumieć*, *wziąć się w garść* oraz *bystry/zaradny*, np.:

(4.1) Boże dziewczyno *ogarnij się*

(4.2) też to jakoś bo normalnie . muszę to napisać jakoś nie wiem . *ogarnąć* to szybko zrobić żeby mieć z głowy

(4.3) to nie ten majk to jest pikuś przy nim . to Kangur jest *ogarnięty* oni się tak nie biją

(4.4) dla mnie to jest nie do *ogarnięcia*. no rozumiem rozumiem Dorota nie każdy

4.2. Typy języka mówionego w NKJP

Mimo niekwestionowanej potrzeby uwzględnienia języka mówionego w dużym korpusie referencyjnym, w zrównoważonej wersji NKJP różnego rodzaju próbki tego typu danych stanowią jedynie około 10 procent korpusu, czyli 30 milionów

[2] Wyjątkiem mogą tu być korpusy bardzo specyficznych odmian angielszczyzny, np. zbiory tekstów piosenek czy prognoz pogody.

Tabela 4.1. Najczęstsze wyrazy w języku mówionym

Nr	Wyraz	Częstość
1.	to	86 960
2.	nie	82 714
3.	no	61 964
4.	i	54 915
5.	w	47 056
6.	że	40 558
7.	się	40 438
8.	tak	39 677
9.	na	37 979
10.	a	29 358

słów tekstowych. Takie proporcje wynikają przede wszystkim z faktu, że pozyskiwanie, transkrypcja oraz anotacja dużych ilości tego typu danych jest jednym z najbardziej pracochłonnych, a przez to kosztownych etapów budowania dużych korpusów. Podobne proporcje między językiem mówionym a pisanym można znaleźć w takich korpusach referencyjnych jak na przykład Brytyjski Korpus Narodowy. Należy przy tym pamiętać, że nie wszystkie dane mówione w NKJP można uznać za język konwersacyjny. Trzy główne typy danych mówionych w NKJP przedstawia tab. 4.2. Dane *mówione medialne* to transkrypcje audycji ra-

Tabela 4.2. Typy języka mówionego w NKJP

Typ	Liczba segmentów wyrazowych
Mówione medialne	900 000
Mówione konwersacyjne	1 900 000
Inne	27 200 000

diowych oraz programów telewizyjnych nagrywane i anotowane specjalnie na potrzeby NKJP. Dane *mówione konwersacyjne* to zebrane również na potrzeby NKJP transkrypcje rozmów Polaków o dość zróżnicowanym profilu wiekowym, w różnym stopniu wykształconych i pochodzących z różnych regionów kraju. Inne dane mówione to głównie skonwertowane na potrzeby NKJP, dostępne publicznie stenogramy posiedzeń Sejmu oraz sejmowych komisji śledczych. Część danych tej ostatniej kategorii należałoby raczej określić mianem *języka czytanego* niż mówionego[3] (np. liczne wystąpienia posłów), chociaż wśród stenogramów

[3] W angielskiej literaturze korpusowej używa się na określenie takich danych terminu *to-be-spoken*.

z posiedzeń komisji śledczych zdarzają się dość spontaniczne pasaże o charakterze konwersacyjnym, które cechuje duży stopień improwizacji.

Niektóre korpusy referencyjne ze względu na trudności związane z pozyskaniem próbek języka konwersacyjnego poprzestają na gotowych transkrypcjach danych medialnych. Przykładem takiego korpusu jest Korpus Współczesnej Angielszczyzny Amerykańskiej (Corpus of Contemporary American English), który zawiera ok. 80 milionów słów języka mówionego w postaci transkrypcji popularnych programów telewizyjnych oraz radiowych. Oczywiście takie dane są same w sobie niezwykle cenne i interesujące, ale należy podkreślić, że język mówiony programów telewizyjnych i radiowych w dość oczywisty i fundamentalny sposób różni się od nieformalnego języka mniej lub bardziej spontanicznych konwersacji między znajomymi, członkami rodziny, a nawet osobami bliżej sobie nieznanymi, ale rozmawiającymi w nieformalnych okolicznościach. Po pierwsze, wypowiedzi uczestników programów telewizyjnych i radiowych podlegają zazwyczaj pewnym rygorom czasowym i tematycznym. Uczestnicy audycji odpowiadają często na konkretne pytania redaktora prowadzącego, którego zadaniem jest czuwanie nad narzuconą z góry spójnością tematyczną danego programu. Dygresje w wypowiedziach uczestników audycji publicystycznych, jeżeli się zdarzają, to traktowane są z reguły jako odstępstwo od formuły programu. Wyjątkiem od tej reguły są pojedyncze audycje. Na przykład audycję „Rozmowy niedokończone", nadawaną w Radiu Maryja oraz Telewizji Trwam, zgodnie z zasygnalizowaną w jej tytule formułą, często cechuje dość duży stopień swobody w dopuszczalnej długości i spójności tematycznej wypowiedzi uczestników pod warunkiem, że pozostają one w zgodzie z pewnym systemem wartości prezentowanym w audycjach nadawanych na antenie tych stacji. Niemały wpływ na język, którym posługują się uczestnicy programów telewizyjnych, ma także sama świadomość obecności w studio i publicznego charakteru utrwalanej lub transmitowanej przez urządzenia rejestrujące rozmowy. Świadomość ta w oczywisty sposób zawęża nie tylko wybory leksykalne i frazeologiczne rozmówców, ale też stosowane przez nich strategie prowadzenia dyskursu oraz wachlarz technik argumentacyjnych. Z punktu widzenia pragmatyki języka można z kolei zauważyć, że odbiorcami komunikatów wypowiadanych publicznie są, poza bezpośrednimi uczestnikami rozmowy, telewidzowie i radiosłuchacze, co ma niemały wpływ na ich treść oraz formę.

Dane mówione medialne zazwyczaj różnią się też od danych mówionych konwersacyjnych formalnością rejestru. Poza programami typu reality show, w danych medialnych rzadko spotyka się wyrazy, frazeologizmy i zbitki słowne typowe dla nieformalnej polszczyzny konwersacyjnej. Ilustrują to ukazane w tab. 4.3 przykłady *zbitek leksykalnych* charakterystycznych dla języka mówionego. *Zbitkami leksykalnymi* (ang. *lexical bundles*) nazywamy w lingwistyce korpusowej najczęściej

Tabela 4.3. Typowe dla polszczyzny mówionej kombinacje segmentów wyrazowych

Zbitka	Przykład
Rejestr formalny i nieformalny	
na przykład	jak rozumiem nie każda służba **na przykład** . współczesne służby jak na przykład jak CBA . nie każde ich postępowanie kończy się aktem oskarżenia w sądzie a tym się powinny...
po prostu	szuka się haków jako metody **po prostu** pokazania . opinii publicznej że ktoś ma te słabe strony i ukrył je . przed wyborcami .
Rejestr formalny	
jest tak że	**jest tak że** no właśnie bo tak krzyczymy chodźcie z nami a nikt za nami nie idzie
wydaje mi się	ale **wydaje mi się** że co innego pan poseł Grupiński powiedział . bardzo proszę
yy yy	ten okres tutaj powojenny yy przysłużył się tej dewastacji tych obiektów natomiast **yy yy** no w tej chwili tutaj to tylko pokazuje jaki jest potencjał
myślę że	**myślę że** to rzeczywiście jest ułomność . polskiej demokracji nie będzie się ważyć kto bardziej kompetentny .
prawda ?	jeśli są prowokacje to trzeba je ujawnić i powiedzieć że one są nielegalne lub legalne . **prawda ?**
Rejestr nieformalny	
nie ?	kartofli nagotowałam tak najwyżej sobie wiesz nawet krupniku tego ugrzeję włożę kartofli bez . **nie ?**
no ale	tak pani też mnie masowała na młode dziewczyny są młode no . **no ale** wszystko takie tam są .
no i	do Kanady pojechała do syna . **no i** tamci tubylcy . zaczęła tym swoim dzieciom bo dwoje bliźniąt tam miała
no bo	**no bo** to jest wasz ten średni pakiet nie ? nie ten najdroższy tylko ten średni

powtarzające się sekwencje wyrazów w tekstach korpusu[4]. Ta prosta technika korpusowa jest często stosowana w analizie dyskursu mówionego (Biber 2004). Choć wiele z najczęstszych sekwencji segmentów wyrazowych nie tworzy spójnych jednostek składniowych ani nawet frazeologicznych, to wygenerowanie ich listy ukazuje kombinacje wyrazów, które odgrywają ważną rolę w budowie wypowiedzi ustnych. Niektóre z nich (np. *na przykład*, *po prostu*) występują bardzo często

[4] Niektóre z podanych tu przykładów to także kombinacje pojedynczych wyrazów i znaków interpunkcyjnych stosowanych w anotacji danych mówionych w NKJP.

w polszczyźnie mówionej bez względu na poziom formalności rozmowy. Są jednak zbitki leksykalne, których prawdopodobieństwo użycia zależy od rejestru języka mówionego. Za szczególnie charakterystyczne dla formalnego języka mówionego można uznać takie kombinacje segmentów wyrazowych, jak *jest tak że*, *myślę że* czy też *wydaje mi się, że*[5]. W rejestrze formalnym mówionym częściej niż w nieformalnych rozmowach zdarzają się wtrącenia sygnalizujące wahanie mówcy (np. *yy yy*), co można wytłumaczyć presją budowania składniejszych, bardziej wyważonych wypowiedzi na antenie radiowej lub telewizyjnej. Z kolei w swobodnych rozmowach w miejsce *prawda?* pojawia się *nie?*, a kombinacje *no* z innymi wyrazami są dużo częstsze niż w wypowiedziach formalnych, np. *no bo*, *no i*, *no ale*.

4.3. Pozyskiwanie danych konwersacyjnych

W odróżnieniu od nagrań medialnych, których dokonuje się w warunkach studyjnych, pozyskanie autentycznych próbek nieformalnego języka mówionego wymagało zastosowania specjalnej metodologii nagrywania i transkrypcji danych. W pierwszej kolejności, kilkanaście specjalnie przeszkolonych w tym zakresie osób zostało wyposażonych w dobrej klasy dyktafony. Osoby te uzyskały zgodę swoich znajomych oraz członków rodzin na nagrywanie ich rozmów w najbliższej przyszłości. Po dokonaniu nagrania osoby zbierające dane uzyskiwały pisemną zgodę na ich wykorzystanie do celów NKJP. Dzięki temu możliwe było zarejestrowanie wielu spontanicznych rozmów, których uczestnicy nie czuli się skrępowani faktem nagrywania ich wypowiedzi. Uzyskane w ten sposób nagrania zostały przetranskrybowane na format pośredni. W transkrypcjach zmieniono lub usunięto imiona i nazwiska, które mogłyby naruszyć prywatność rozmówców. Osoby nagrywające dane zajmowały się następnie ich transkrypcją, dzięki czemu łatwiej było odtworzyć fragmenty rozmów zarejestrowanych w trudnych warunkach akustycznych.

4.4. Anotacja

Schemat anotacji używany do transkrypcji danych mówionych różni się dość istotnie od anotacji tekstów pisanych. Każda nagrana rozmowa jest traktowana jako jeden tekst, w którym podstawową jednostką strukturalną są kolejne wypowiedzi mówców. Przykład formatu anotacji języka mówionego używanego do transkrypcji rozmów w NKJP przedstawiono na wydr. 4.1. Warto zaznaczyć, że jest to schemat anotacji zgodny ze standardem TEI (Text Encoding Initiative).

[5] Typowość kombinacji dla danego rejestru zmierzono miarą istotności statystycznej przez porównanie częstości względnych w danych mówionych medialnych oraz konwersacyjnych.

Wydruk 4.1. Przykład anotacji strukturalnej z podziałem na wypowiedzi

```
1  <u who="sp3" xml:id="u-1"><incident><desc>pukanie do
   drzwi</desc></incident> proszę. bileciki do kontroli.</u>
2  <u who="sp2" trans="overlap" xml:id="u-2"> co tam ?</u>
3  <u who="sp0" trans="overlap" xml:id="u-3"> cześć cześć
   cześć.</u>
4  <u who="sp6" xml:id="u-4"> czemu wy stoicie ?</u>
5  <u who="sp2" xml:id="u-5"> siadajcie rozgoście się impreza
   się tu przeniosła. czekacie na pocią ?<gap/> jak w
   poczekalni zupełnie. jak na pociąg byśmy czekali.</u>
6  <u who="sp3" xml:id="u-6"> jak na terminalu.</u>
7  <u who="sp2" xml:id="u-7"> na terminalu. ta
   Londyńczycy.<vocal> <desc>śmiech</desc> </vocal></u>
8  <u who="sp4" xml:id="u-8"> dziewczyny a wy już spakowane
   jesteście ?</u>
9  <u who="sp0" xml:id="u-9"> no ta. miałam tyle rzeczy że ho.
   pół szafy zabrałam. co ?</u>
10 <u who="sp2" xml:id="u-10"> nie ma to jak na piecyku nie
   ?</u>
11 <u who="sp1"
   xml:id="u-11"><vocal><desc>ziewając</desc></vocal>
   nie.<vocal><desc>#VOICE_END</desc></vocal> on nie grzeje
   wiesz.</u>
12 <u who="sp0" xml:id="u-12"> Darek. my musimy kluczyki od
   razu oddawać czy to jak to wygląda ? czy do ciebie ?</u>
```

Poszczególne wypowiedzi przypisane są do rozmówców zdefiniowanych w nagłówku pliku transkrypcji. Dzięki temu możliwe jest przeszukiwanie pełnotekstowe wypowiedzi z uwzględnieniem podstawowych kryteriów socjolingwistycznych, takich jak wiek, wykształcenie, płeć czy też miejsce zamieszkania mówców. W transkrypcji języka mówionego wielkie litery używane są tylko w imionach oraz dobrze znanych nazwach własnych. Obce nazwy i wyrazy zapisywane są fonetycznie. Fonetycznie zapisywana jest także wyraźna wymowa gwarowa lub niestandardowa, a także przypadki stylizacji na taką wymowę. Wykrzyknik oraz pytajnik pełnią funkcję podobną do ich odpowiedników w standardowej interpunkcji. Kropki są stosowane do oznakowania pauz w pojedynczych wypowiedziach i nie można ich uważać za granice zdań. Istotne zmiany tonu głosu oraz zdarzenia mogące mieć wpływ na przebieg rozmowy zaznaczono specjalnymi elementami `<vocal>` oraz `<incident>`. Fragmenty rozmów, których ze względu na warunki akustyczne nie udało się przetranskrybować, zostały zastąpione znacznikiem `<gap>`. Nakładanie się na siebie wypowiedzi sygnalizuje

wartość atrybutu `@trans="overlap"`. W najbliższej przyszłości planowane są również prace nad dodaniem do wybranych transkrypcji anotacji czasowej, która umożliwi swobodne odtwarzanie dowolnego fragmentu rozmowy.

Anotacja zdarzeń ma niekiedy szczególne znaczenie ze względu na zubożenie semiotyczne, jakie cechuje zapis ortograficzny autentycznych rozmów, toczonych często w warunkach „polowych". Przykładem takiej sytuacji może być fragment rozmowy ukazany na wydr. 4.2.

Wydruk 4.2. Anotacja zdarzeń w transkrypcjach NKJP

```
1 <u who="sp2" trans="overlap" xml:id="u-105"> to i to i to i
  tak taniej a taki kupić sobie pięcio– sześcioletni rozumiesz
  audi czy to jest samochód do śmierci. to takie tutaj
  pokupowali audiki można powiedzieć dziesięcio–
  piętnastoletnie jak ten Szymanek kupił dizla . też
  sprowadzany też mu sprowadził Niemiec nie ? kurde to ci
  mówię samochód nie do zdarcia. i hak i przyczepę widziałeś
  przecież. a nie byłeś tam<gap/><incident><desc>wszyscy
  spoglądają na kaczkę która je kapselek od
  piwa</desc></incident></u>
2 <u who="sp0" xml:id="u-106"> kapselek je !</u>
3 <u who="sp2" xml:id="u-107"> no tak to wszystko tylko żeby
  nie pożarł. kaczka !<vocal><desc>śmiech</desc></vocal>
  widzisz jakie kaczory to głupki</u>
4 <u who="sp0" xml:id="u-108"> o drugi o</u>
5 <u who="sp2" xml:id="u-109"> kiedyś wyciągłem to ze dwa
  metry sznurka</u>
6 <u who="sp0" trans="overlap" xml:id="u-110"> o Jezu</u>
7 <u who="sp2" xml:id="u-111"> to taki
  duży<incident><desc>pokazuje jaki długi
  sznurek</desc></incident> znalazł sznurek rozumiesz i tak
  połykał połykał połykał i później przestała żreć ta kaczka i
  se latała z tym sznurkiem i czubek było widać mówię ciągnę
  ciągnę<vocal><desc>śmiech</desc></vocal>
```

Tematem rozmowy toczonej upalnego lipcowego dnia w gospodarstwie rolnym w województwie kujawsko-pomorskim są sprowadzane z zagranicy samochody. W pewnej chwili, dość niespodziewanie, w transkrypcji pojawia się wypowiedź *Kapselek je!* Didaskalia dodane przez osobę transkrybującą tekst w znaczniku `<incident>` pomagają zrozumieć tę nagłą zmianę tematu, którą w tym wypadku wywołała kaczka próbująca połknąć leżący na podwórzu kapsel od butelki.

W nagłówkach trankrypcji języka mówionego znajdują się informacje o poszczególnych uczestnikach rozmowy (zob. wydruk 4.3), przy czym w transkrypcjach audycji radiowych i telewizyjnych brakuje czasem informacji o miejscu zamieszkania, wieku oraz wykształceniu niektórych rozmówców. Dodatkowo w nagłówkach zakodowano miejsce i datę dokonania nagrania oraz główne tematy poruszane podczas rozmowy. Co ciekawe, z anotacji tematów rozmów wynika, że różnorodność tematyczna jest jedną z charakterystycznych cech potocznego języka konwersacyjnego i nie zawsze musi być bezpośrednim rezultatem zdarzeń występujących w trakcie samej rozmowy. W ciągu godzinnej rozmowy znajomych często poruszonych zostaje kilkadziesiąt tematów. Niektóre z nich są porzucane tuż po wprowadzeniu, choć zdarzają się także dłuższe okresy stabilizacji tematycznej.

Wydruk 4.3. Nagłówek transkrypcji języka konwersacyjnego

```
<particDesc>
 <person xml:id="sp1" role="speaker">
   <persName>anonim</persName>
   <education xml:lang="pl">wyższe</education>
   <age>25</age>
   <residence>Wieluń</residence>
 </person>
 <person xml:id="sp0" role="speaker">
   <persName>anonim</persName>
   <sex value="1">male</sex>
   <education xml:lang="pl">średnie</education>
   <age>49</age>
   <residence>Wieluń</residence>
 </person>
</particDesc>
```

4.5. Wyszukiwarka

W oparciu o wspomniany powyżej schemat anotacji opracowana została specjalna wyszukiwarka pozwalająca wygodnie przeszukiwać, sortować i wyświetlać konwersacyjną część korpusu NKJP z uwzględnieniem metadanych właściwych dla języka mówionego. Rys. 4.1 ukazuje przykładowy ekran kontekstu wyniku pasującego do zapytania `no i`. Poszczególne wypowiedzi stanowiące kontekst wystąpienia poszukiwanego wyrazu lub frazy ukazane są w kolejnych wierszach tabeli. W ostatnich trzech kolumnach wyświetlane są informacje o płci,

wieku i wykształceniu autora danej wypowiedzi. Wyszukiwarka NKJP dla danych mówionych jest dostępna publicznie pod adresem `http://nkjp.uni.lodz.pl/spoken.jsp`.

Rysunek 4.1. Ekran wyników w wyszukiwarce danych konwersacyjnych

Informacje o nagraniu

Id	Tytuł	Źródło	Data nagrania
31	*Motor*	PELCRA	2008

Kontekst wystąpienia

#	Id	Wypowiedź	Płeć	Wiek	Wykształcenie
1.	3616	czy słono ? normalnie wiesz tak jak	m	34	wy
2.	3617	normalne ceny . ale	m	38	wy
3.	3618	kupił to tam za czterysta tysięcy bodajże	m	34	wy
4.	3619	to faktycznie jak za taką działę	m	38	wy
5.	3620	w tamtym miejscu wiesz chata dwieście osiemdziesiąt metrów czy tam blisko dwustu osiemdziesięciu	m	34	wy
6.	3621	i dwa poziomy ?	m	38	wy
7.	3622	*ogromna działka . dwa poziomy mieszkalne* ***no i*** *piwnica taka wiesz wielka tak że*	m	34	wy
8.	3623	kawał domu	m	38	wy
9.	3624	kawał domu . no i wszystko później mu robili . i kazał wiesz . bo oczywiście robotnicy chcieli już tam schody rozwalać . tu te a schody też fajne drewniane takie pół kręcone jakby na piętro nie ? tu już to . wszystko wiesz . zero rozwalania mają wszystko wiesz	m	34	wy

4.6. Przyszłe prace

Należy podkreślić, że zebranie tak dużej liczby transkrypcji języka potocznego było trudnym zadaniem i wymagało pewnych kompromisów. Z praktyki zbierania tego typu danych wynika, że ze względu na zakłócenia i warunki akustyczne, danej rozmowy często nie może w całości odtworzyć osoba, która nie brała w niej choćby biernego udziału. Jednakże nie było ani możliwe, ani metodologicznie uzasadnione powierzenie zbierania danych konwersacyjnych tylko głęboko uświadomionym językoznawcom lub osobom o nienagannej ortografii, ponieważ z pewnością obniżyłoby to poziom zróżnicowania zebranej próbki. W transkrypcjach danych konwersacyjnych można więc czasem spotkać niejednolity zapis niestandardowej, gwarowej lub stylizowanej wymowy, które mogą wymagać dalszej normalizacji w kontekście całego korpusu. Sporadycznie pojawiają się także błędy ortograficzne, choć często trudno bezpośrednio przekładać normy ortografii języka pisanego na zapis swobodnych rozmów[6]. Zebranie

[6] Co ciekawe, oczywiste błędy ortograficzne występują także w transkrypcjach mówionych Brytyjskiego Korpus Narodowego. Mylone są na przykład formy *their* i *they're*.

tak dużej liczby stosunkowo zróżnicowanych nagrań swobodnych rozmów oraz przetranskrybowanie ich w stosunkowo krótkim czasie rządziło się swoimi prawami. Natomiast dzięki zachowanym nagraniom cyfrowym korekta ortografii i normalizacja transkrypcji może być przedmiotem dalszych prac stosunkowo niewielkiej liczby osób, które nie musiały brać udziału w nagranych rozmowach.

4.7. Podsumowanie

Nie ulega wątpliwości, że autentyczny język konwersacyjny istotnie różni się od innych odmian polszczyzny. Z punktu widzenia badacza, dyskurs konwersacyjny cechuje pewna nieprzewidywalność, która częściowo wynika z przebiegu samej rozmowy, a częściowo także z okoliczności, w których jest ona prowadzona, i zdarzeń, które jej towarzyszą. Zamiast kompletnych składniowo wypowiedzi, które znamy z języka pisanego, w języku konwersacyjnym często mamy do czynienia z aproksymacją komunikowanych znaczeń (Lewandowska-Tomaszczyk 2012). Trudno jest w autentycznej polszczyźnie konwersacyjnej wskazać granice zdań składniowych, a w warstwie leksykalnej szczególnie często występują wyrazy, frazy i zbitki leksykalne, które pełnią funkcję dyskursywną i ułatwiają budowanie wypowiedzi ustnych w czasie rzeczywistym. Podstawowa funkcja korpusu referencyjnego, jakim jest Narodowy Korpus Języka Polskiego, to reprezentacja poczucia językowego przeciętnego użytkownika polszczyzny. Aby korpus spełniał tę funkcję, konieczne było zebranie możliwie największej liczby próbek języka mówionego, ze szczególnym uwzględnieniem potocznego języka konwersacyjnego. Według wiedzy autora, komponent konwersacyjny NKJP jest największym polskim korpusem języka potocznego o dużym stopniu zróżnicowania sytuacyjnego. Pozostaje mieć nadzieję, że będzie on wykorzystywany w ilościowej oraz jakościowej analizie polszczyzny mówionej.

Część III

Zasady znakowania

5

Ręcznie znakowany milionowy podkorpus NKJP

Łukasz Degórski, Adam Przepiórkowski

5.1. Idea

Najbardziej pracochłonnym zadaniem w projekcie NKJP było stworzenie i ręczne oznakowanie zrównoważonego (w sensie sprecyzowanym w rozdz. 3) podkorpusu składającego się z niewielkich próbek różnorodnych tekstów. Podkorpus ten liczy nieco ponad milion tradycyjnie rozumianych słów (a zatem istotnie ponad milion segmentów), stąd posługujemy się w niniejszej publikacji skrótowymi terminami „(pod)korpus milionowy" i „(pod)korpus 1M".

Do idei znakowania korpusów nikogo przekonywać nie trzeba: oprócz zrównoważenia, to właśnie obecność dodatkowych informacji o tekstach, w tym informacji lingwistycznych, jest jedną z głównych przyczyn tego, że nadal konstruuje się korpusy, zamiast zadowolić się znacznie przecież większymi zasobami Internetu. Standardowym poziomem lingwistycznym, obecnym w prawie każdym korpusie godnym tego miana, jest poziom morfosyntaktyczny, na którym każde słowo ma przypisaną mu informację o formie podstawowej, części mowy i wartościach choćby podstawowych kategorii morfologicznych, takich jak rodzaj i przypadek. Coraz więcej jest jednak korpusów znacznie wychodzących poza ten poziom, w tym – przede wszystkim – tzw. banków drzew (ang. *treebanks*), a więc korpusów anotowanych składniowo.

Aby osiągnąć jak najwyższą jakość znakowania, należałoby je przeprowadzać nie automatycznie, lecz ręcznie i w dodatku z zachowaniem dosyć rygorystycznej procedury znakowania: każdy fragment tekstu powinien być niezależnie anotowany przez dwoje lingwistów, a wszelkie rozbieżności powinny być rozstrzygane przez szczególnie doświadczonego anotatora, zwanego czasem potocznie

superanotatorem (por. p. 6.6, p. 8.4 i p. 13.2). Oczywiście taka procedura jest bardzo kosztowna, można ją więc stosować wyłącznie do korpusów o niewielkich rozmiarach, zwykle nieprzekraczających miliona słów.

W wypadku korpusów znacznie większych, takich jak należący obecnie do największych na świecie Narodowy Korpus Języka Polskiego, jedyną możliwością jest anotacja automatyczna, za pomocą specjalnie do tego stworzonych programów komputerowych. Programy takie muszą się jednak nauczyć znakować teksty, a do tego potrzebują… tekstów znakowanych przez człowieka[1]. Stąd właśnie potrzeba stworzenia ręcznie anotowanego podkorpusu NKJP.

W niniejszym projekcie przyjęto następujące poziomy znakowania korpusu:

1. segmentacja tekstu na zdania i słowa (a właściwie: segmenty wyrazowe),
2. znakowanie morfosyntaktyczne,
3. znakowanie sensami słów (w ograniczonym zakresie),
4. powierzchniowe znakowanie składniowe, tj. anotacja:
   a) słów składniowych i
   b) prostych grup składniowych,
5. znakowanie jednostek nazewniczych.

W innych projektach realizowanych w IPI PAN przedstawiony tutaj podkorpus milionowy jest ponadto znakowany pełnymi drzewami składniowymi.

Zasady ręcznego znakowania na różnych poziomach lingwistycznych zostały szczegółowo omówione w:

– rozdz. 6 – segmentacja i znakowanie morfosyntaktyczne,
– rozdz. 7 – wybór ponad stu leksemów i ich znaczeń do znakowania sensami słów,
– rozdz. 8 – powierzchniowe znakowanie składniowe,
– rozdz. 9 – znakowanie jednostek nazewniczych.

Niełatwym zadaniem okazało się stworzenie schematu reprezentacji tych wszystkich informacji – poświęcony temu został raczej techniczny rozdz. 10.

Także w ramach projektu NKJP stworzone zostały narzędzia do automatycznego znakowania pełnego korpusu, w tym czy innym sensie „wytrenowane" na korpusie milionowym. Tak więc tzw. tager morfosyntaktyczny został omówiony w rozdz. 11, system do semantycznego ujednoznaczniania wybranych leksemów – w rozdz. 12, a narzędzie do znakowania jednostek nazewniczych – w rozdz. 13. (Osobnego rozdziału nie poświęcono poziomowi powierzchniowo-składniowemu, gdyż do oznakowania całego NKJP posłużyła zmodyfikowana gramatyka opisana w rozdz. 8). Zauważmy jednak, że – choć cały NKJP został

[1] Bezpośrednie sformalizowanie lingwistycznych reguł znakowania okazuje się zadaniem znacznie trudniejszym niż stworzenie takiego „korpusu treningowego".

oznakowany na wszystkich tych poziomach – obecne wersje wyszukiwarek obsługujących pełny korpus oferują jedynie dostęp do poziomu morfosyntaktycznego[2]. Mamy nadzieję, że przyszłe projekty pozwolą na skonstruowanie interfejsu do pozostałych poziomów znakowania NKJP.

W dalszej części rozdziału opisujemy próbkowanie tekstów do podkorpusu milionowego (p. 5.2) oraz wspominamy o zasadach jego udostępniania (p. 5.3).

5.2. Konstrukcja podkorpusu

5.2.1. Założenia

Dostępne w momencie tworzenia podkorpusu milionowego teksty korpusu NKJP zostały podzielone według stylu i źródła na kategorie podane w tab. 5.1.

Tabela 5.1. Kategorie tekstów

Kategoria	Udział w korpusie
Dzienniki	25,5%
Pozostałe periodyki	23,5%
Książki publicystyczne	1,0%
Literatura piękna	16,0%
Literatura faktu	5,5%
Typ informacyjno-poradnikowy	5,5%
Typ naukowo-dydaktyczny	2,0%
Internetowe interaktywne (blogi, fora, Usenet)	3,5%
Internetowe nieinteraktywne (statyczne strony, Wikipedia)	3,5%
Quasi-mówione (protokoły sesji parlamentu)	2,5%
Mówione medialne	2,5%
Mówione konwersacyjne	5,0%
Inne teksty pisane	3,0%
Książka niebeletrystyczna nieklasyfikowana	1,0%

Wykorzystane zostały wyłącznie teksty współczesne (stworzone po 1945 roku).

Podkorpus milionowy miał spełniać następujące założenia:

1. Powinien być zrównoważony w sensie kategorii tekstów według podanych wyżej proporcji.

[2] Ściślej rzecz ujmując, informacje te są widoczne – i mogą być użyte jako kryteria wyszukiwania – w stworzonej w IPI PAN wyszukiwarce Poliqarp (Janus i Przepiórkowski 2007; `http://poliqarp.sourceforge.net/`; `http://nkjp.pl/poliqarp`.). Druga wyszukiwarka, opisana szczegółowo w rozdz. 14, pozwala natomiast na wyszukiwanie fleksyjne słownikowe, czerpiąc informacje morfosyntaktyczne ze słownika Morfologik (`http://morfologik.blogspot.com/`).

2. Powinien być złożony z krótkich próbek poszczególnych tekstów, przy czym próbki powinny być pobierane z losowego miejsca tekstu źródłowego (a więc nie akceptujemy np. algorytmu, który po prostu brałby początek każdego tekstu).
3. Pojedyncza próbka powinna stanowić logiczną całość, w miarę możliwości całe zdania (na przykład nie powinna składać się z końca jednego i początku drugiego akapitu ani urywać w połowie zdania złożonego), z przewagą rzeczywistego materiału językowego – bez nadmiaru cyfr, symboli itp.
4. Powinna być użyta tylko bardzo niewielka część każdego tekstu źródłowego.
5. W miarę możliwości każdy tekst wejściowy jest reprezentowany w podkorpusie, a jeśli to niemożliwe – wybór powinien być losowy, a nie arbitralny.

5.2.2. Algorytm

Rozmiar pojedynczej próbki ustalony został na 40 do 70 słów. Intuicyjnie odpowiada to kilku zdaniom (średnia długość polskiego zdania to około 16 słów), a przy tym zapewnia porównywalność z próbkami w korpusie Słownika Frekwencyjnego. W toku prac, na skutek drobnych modyfikacji i poprawiania błędów, pojedyncze (bardzo nieliczne) próbki mogły zostać skrócone nieco poniżej długości 40 słów.

Istnieje teoretyczna możliwość, że wylosowane zostały próbki, które w oryginalnym tekście sąsiadowały ze sobą.

Wejściowe teksty zostały podzielone według kategorii. Dla każdej kategorii na podstawie liczby słów w dostępnych tekstach tej kategorii dobrany został współczynnik próbkowania taki, by wylosowana została w sumie żądana liczba słów. Wiązały się z tym następujące problemy:

1. W niektórych kategoriach, np. dziennikach, pojedyncze teksty źródłowe były bardzo małe, rzędu jednego krótkiego artykułu lub notki, a współczynnik próbkowania na tyle niski, że nie byłoby szans nawet na jedną dłuższą niż 40 słów próbkę z każdego tekstu.
2. W niektórych kategoriach było odwrotnie: materiału było na tyle mało, że współczynnik przekraczał 5%, co utrudniało losowanie, zwłaszcza jeśli złożyło się to z techniczną niedoskonałością tekstów źródłowych.
3. W pewnym podzbiorze tekstów internetowych materiał składał się zarówno z bardzo małych, jak też bardzo dużych plików; dobranie jednego współczynnika spowodowałoby, że duże teksty zupełnie zdominowałyby próbkę w tej kategorii, co byłoby sprzeczne z założeniem 5.

Problem 1 został rozwiązany metodą łączenia mniejszych tekstów w agregaty. Agregowane były bądź całe tytuły (wszystkie teksty z danego czasopisma znalazły się w jednym pliku), bądź gdy materiału było więcej – roczniki. Posłużenie się agregacją spowodowało, że nie każdy tekst wejściowy jest reprezentowany w milionowym podkorpusie, ale pozostało w zgodzie z założeniem 5: każdy tekst miał teoretyczną szansę się tam znaleźć. Wspomniany podział na roczniki pozwolił zachować rozsądną wielkość agregatów, a przy tym zapewnił, że teksty z każdego rocznika znajdą się w losowanym podkorpusie.

Problem 2 wydaje się mało poważny, ale w związku z technicznymi wadami tekstów oraz dodatkowymi ograniczeniami nałożonymi na kształt próbki zdarzało się, że wylosowanie próbek stanowiących około 5% tekstu źródłowego było trudne – jak się okazało, założenie 3 powoduje, że w niektórych tekstach, np. naukowo-dydaktycznych, sporo wylosowanych próbek musiało być uznane za niepoprawne. Szczególnie gdy duża część akapitów okazywała się wadliwa z powodów technicznych, np. kończyły się w połowie zdania. Takie akapity były ignorowane bądź – wybrane – ręcznie naprawiane, aby dało się wylosować odpowiednią liczbę próbek. W takim przypadku w pewnym stopniu zakłócana była losowość wyboru – ręcznie naprawiano tylko mniej więcej tyle akapitów, by wystarczyło, a więc de facto losowania dokonywała osoba poprawiająca akapity. Jednakże, by zminimalizować to zakłócenie, ręcznego wyboru akapitów do poprawienia staraliśmy się również dokonywać losowo.

Problem 3 udało się rozwiązać, dzieląc problematyczne teksty na 27 podklas według wielkości, i przydzielając odpowiednią docelową liczbę słów każdej podklasie. Doświadczalnie dobrany został algorytm, w którym liczba słów w próbce z każdej podklasy jest proporcjonalna do pierwiastka sześciennego z liczności tej podklasy mierzonej w słowach tekstu źródłowego.

Trudne okazało się spełnienie założenia 3, po części dlatego, że sformułowane jest ono z konieczności dość nieprecyzyjnie. Na pierwszym etapie odrzucane były akapity krótsze niż 25 słów, które nie wyglądają jak zdanie, oraz te akapity dłuższe, które nie rozpoczynają się jak zdanie (kończyć się mogą dowolnie). Kryterium „wyglądania jak zdanie" to w uproszczeniu wielka litera na początku i odpowiedni znak przestankowy na końcu. To jednak nie wystarczyło. Aby wyeliminować znaczną grupę próbek niestanowiących – mimo „wyglądania jak zdania" – spójnego materiału językowego, wprowadzona została wymagana średnia długość akapitu: więcej niż pięć słów. Oznacza to, że próbka nie może być np. ciągiem dwunastu czterosłowowych akapitów, ale nadal może w niej wystąpić akapit nawet jednosłowowy, jeśli tylko inne akapity w tej samej próbce będą na tyle długie, by zwiększyć średnią powyżej pięciu. To wyeliminowało np. próbki, które były listą wyników meczów piłkarskich.

Mimo zastosowania wszystkich wymienionych procedur, po wygenerowaniu próbek okazało się konieczne dodanie kolejnego etapu przygotowania – ręcznego sczytywania. Tak wielka masa tekstów wejściowych, pochodzących z różnych źródeł, zawierała zbyt dużo nietypowych błędów – formatowania, podziału na elementy, literówek typu technicznego – i algorytmy automatycznie nie mogły sobie z nimi wszystkimi wystarczająco dobrze poradzić. W trosce więc o jakość podkorpusu, przeznaczonego wszak do bardzo pracochłonnej i wymagającej dobrego materiału ręcznej anotacji na wielu poziomach, każda wylosowana próbka była czytana i ewentualnie usuwana lub skracana. Proporcje podkorpusu nie zostały przy tym zakłócone – usunięte próbki były zastępowane ponownie wylosowanymi z tych samych tekstów, a od pewnego momentu generowane były z pewnym zapasem (który należało dla równowagi usunąć, jeśli większość okazała się poprawna).

5.2.3. Implementacja

Wszystkie akapity tekstów wejściowych, jako elementy XML, opatrywane są informacją, z jakiego miejsca którego pliku pochodzą. To umożliwi ich identyfikację, gdy zostaną oderwane od plików źródłowych. Informacja taka umożliwia zabezpieczenie się przed wylosowaniem próbki, która w połowie składa się z końcówki jednego tekstu, a w połowie – z początku kolejnego; ma to znaczenie w wypadku agregatów. Następnie liczona jest jego długość każdego akapitu w słowach. Za słowo uznawany jest dowolny ciąg znaków między białymi znakami (`\s+`).

Jako nieużyteczne oznaczane są akapity:

1. o zerowej długości,
2. o długości mniejszej niż 25 słów, które nie pasują do wyrażenia regularnego:
   `/^\s*[\(\"]?(\N{EN DASH}\s|\s)?([0-9A-Z]|Ą|Ć|Ę|Ł|Ó|Ń|Ś|Ź|Ż).*`
   `[\.!?\x{2026}\x{2025}:]+[\"\)]*\s*$/,`
3. dłuższe, które nie pasują do wyrażenia:
   `/^\s*[\(\"]?(\N{EN DASH}\s|\s)?([0-9A-Z]|Ą|Ć|Ę|Ł|Ó|Ń|Ś|Ź|Ż).*$/.`

Oznaczenie odbywa się przez sztuczne dodanie 9000 do obliczonej długości akapitu. W takim przypadku staje się on za długi i na pewno nie zostanie potem wylosowany do próbki. Nie można go po prostu usunąć, bo z ciągu `...aXc...`, gdzie `X` jest tym odrzuconym akapitem, dałoby się potem wylosować próbkę `ac`, która nie byłaby ciągłym podzbiorem tekstu oryginalnego.

Przez tak przygotowaną listę akapitów przechodzimy następnie, korzystając z dwóch znaczników – pozycji początku i końca. Posługiwać się będziemy poniżej trzema terminami:

1. *akapit* (z obliczoną długością),
2. *potencjalna próbka* – ciągła grupa akapitów, spełniająca warunki nakładane na próbki, takie jak minimalna i maksymalna długość itp.,
3. *kandydat* (na potencjalną próbkę) – dowolna ciągła grupa akapitów, wyznaczona znacznikiem początku i końca na liście akapitów.

Oba znaczniki ustawiamy na 1. W każdym momencie znamy długość aktualnego kandydata w słowach (suma długości akapitów składowych) oraz średnią długość akapitu. Jeśli długość jest zbyt duża, zwiększamy o 1 znacznik początku (i jednocześnie znacznik końca, jeśli były sobie równe). Jeśli długość jest za mała, zwiększamy o 1 znacznik końca, otrzymując dłuższego kandydata. Jeśli długość jest prawidłowa, przetwarzamy kandydata i przesuwamy o 1 znacznik końca. Przetworzenie kandydata polega na sprawdzeniu, czy średnia długość akapitu jest większa niż 5. Jeśli tak, kandydat awansuje na potencjalną próbkę – we wszystkich akapitach składowych dopisywana jest informacja, że są one jej elementami. W rezultacie każdy akapit może być elementem dowolnej liczby potencjalnych próbek.

Na kolejnym etapie następuje właściwe losowanie wymaganej liczby próbek. Przebiega ono następująco:

1. Z listy akapitów wybierany jest losowo jeden.
2. Jeśli jest on elementem jakichś potencjalnych próbek, wybierana jest z nich ta, której długość najbliższa jest 55 słowom, i staje się ona ostateczną próbką. Jako taka jest wypisywana w docelowym formacie i oznaczona (wraz ze swoimi akapitami składowymi) jako użyta. Potencjalna próbka, która zawiera choć jeden użyty wcześniej akapit, nie będzie więcej brana pod uwagę.
3. Operacja powyższa jest powtarzana do czasu przekroczenia zadanej liczby słów we wszystkich próbkach. Ostatnia (powodująca przekroczenie) potencjalna próbka jest dodawana lub nie, w zależności od tego, czy zwiększyłaby, czy zmniejszyła wartość bezwzględną różnicy między zadaną a otrzymaną liczbą słów. Intuicyjnie oznacza to, że lepiej np. wziąć o 30 za mało, niż o 31 za dużo.

5.3. Dostępność

Z powodu praw autorskich nie jest możliwe – nawet niekomercyjne – udostępnianie pełnego NKJP w postaci tekstowej. Inaczej ma się sprawa w wypadku podkorpusu milionowego, składającego się z krótkich fragmentów większych tekstów, w dodatku użytych jako cytaty, będące podstawą znacznie istotniejszych od samych tekstów informacji lingwistycznych. Korpus ten – wraz ze wszystkimi

poziomami anotacji – udostępniany jest na swobodnej licencji GNU GPL (wersja 3) w postaci XML-owej opisanej w rozdz. 10. Udostępniając ten podkorpus in extenso, liczymy na to, że zostanie wykorzystany przez społeczność lingwistyczną i informatyczną, m.in. do trenowania narzędzi do przetwarzania języka naturalnego i rozszerzony o dalsze poziomy znakowania lingwistycznego[3]. W chwili pisania niniejszego artykułu dostępna jest – na stronach *Computational Linguistics in Poland* (`http://clip.ipipan.waw.pl/LRT`) – pierwsza wersja podkorpusu milionowego; sukcesywnie będą się na niej pojawiały także wersje kolejne.

[3] Planowane jest m.in. dodanie poziomu nawiązań (tj. konstrukcji koreferencyjnych) w ramach projektu CORE (*Komputerowe metody identyfikacji nawiązań w tekstach polskich*; `http://clip.ipipan.waw.pl/CORE`).

6

Anotacja morfoskładniowa

Łukasz Szałkiewicz, Adam Przepiórkowski

6.1. Wstęp

Do ręcznej anotacji morfoskładniowej z całego Narodowego Korpusu Języka Polskiego wybrano podkorpus liczący milion słów (zob. rozdz. 5). Anotacja ta oznaczała:

1. podział:
   a) na zdania,
   b) na segmenty,
2. lematyzację segmentów,
3. przypisanie segmentom „części mowy”[1],
4. przypisanie segmentom wartości kategorii gramatycznych.

Ostatnie trzy z wymienionych zadań polegały na przypisaniu segmentowi odpowiedniego znacznika (tagu). Każdy znacznik zawierał formę podstawową (hasłową), wskazującą na odpowiedni leksem, oraz podaną jednym ciągiem po dwukropkach „część mowy” i wartości odpowiednich kategorii gramatycznych. Przykładowy tag segmentu *słowników* we fragmencie *rynek słowników pedagogicznych* to SŁOWNIK subst:pl:gen:m3 (lub słownik:subst:pl:gen:m3). Czasem będziemy też używać słowa *znacznik* (*tag*) w węższym rozumieniu, jako odnoszącego się do opisu niezawierającego formy podstawowej (czyli ograniczonego do subst:pl:gen:m3). Zapis wybranych wartości oddzielonych kropką oznaczać będzie ich alternatywę, np. subst:sg:nom.acc.voc:n to rzeczownik rodzaju nijakiego w liczbie pojedynczej w jednym z przypadków: mianownik, biernik, wołacz. Zbiór wszystkich znaczników nazywamy *tagsetem*.

[1] „Części mowy” ujmujemy tu w cudzysłów, gdyż w NKJP przyjęto dużo bardziej szczegółowy podział na fleksemy (zob. p. 6.3.1).

W pierwszych punktach niniejszego rozdziału (p. 6.2–p. 6.5) skupimy się na lingwistycznych aspektach ręcznej anotacji morfoskładniowej podkorpusu milionowego, a jej robocze oraz techniczne aspekty zostaną omówione w p. 6.6. Narzędzie do automatycznej anotacji pełnego korpusu znacznikami morfosyntaktycznymi zostanie natomiast opisane w rozdz. 11.

Trzeba pamiętać, że ręczna anotacja morfoskładniowa przynosiła każdego dnia nowe problemy językowe. Choć istniała odpowiednia instrukcja znakowania, to wymagała ona czasem drobnych korekt, a stale – uzupełniania. Niestety, dla zachowania spójności anotacji na pewnym etapie należało zaprzestać zmian w instrukcji. Niektóre rozwiązania były więc z pewnością kompromisem między chęcią analizowania każdej kwestii tak szczegółowo, jak czasem by się tego pragnęło, a ograniczonym czasem i środkami na wykonanie zadania.

Lingwistyczne zasady znakowania NKJP wywodzą się wprost z zasad znakowania ustalonych w trakcie prac przy tagowaniu Korpusu IPI PAN[2]. Część rozstrzygnięć została jednak zmieniona, inne przemyślane i doprecyzowane. Przy znakowaniu Korpusu IPI PAN był wykorzystywany analizator morfologiczny Morfeusz SIaT (Woliński 2006). Również w anotacji NKJP użyto Morfeusza, z tym że w innej wersji (Morfeusz SGJP), opartego na danych ze *Słownika gramatycznego języka polskiego* (Saloni i in. 2007). Tenże program jest obecnie dostępny na licencji BSD na stronie `http://sgjp.pl/morfeusz/`.

Wykorzystanie Morfeusza nie oznaczało jednak znakowania tekstu bezpośrednio za jego pomocą, lecz analizę słów przez podawanie anotatorom do wyboru wszystkich możliwych interpretacji danego segmentu w tekście (więcej o tym w p. 6.6.3). Oczywiście informacje programu były rzutowane na przyjęty w projekcie tagset (np. liczba rodzajów musiała zostać zredukowana z dziewięciu do pięciu).

6.2. Segmentacja

6.2.1. Podział na zdania

Zdanie w anotacji morfosyntaktycznej rozumiane było tradycyjnie, jako ciąg od wielkiej litery do kropki lub innego znaku kończącego (wielokropek, pytajnik, wykrzyknik). W razie wątpliwości uznawano za zdanie ciąg dłuższy, np. jednym zdaniem były:

[2] Stworzony w latach 2001–2004 w ramach grantu KBN i rozwijany potem w ramach prac własnych IPI PAN; wszelkie szczegóły wraz z bogatą dokumentacją w postaci licznych publikacji – w tym podsumowującą publikację Przepiórkowski 2004 – można znaleźć na stronie: `http://korpus.pl/`.

(6.1) *Czytanie tego tekstu ubogaca – rzekła Maria.*
(6.2) *– Jak bardzo mógł się zmienić, pokazał wczoraj, gdy krzyczał do nas: „Niewiele da się tu poprawić, nie widzicie tego."* (znak końca na ostatnim cudzysłowie)

6.2.2. Podział na segmenty

Segmenty nie mogły być dłuższe niż *słowa* rozumiane jako maksymalne ciągi znaków nie będących separatorami słów. Za *separatory słów* uznano odstępy (spacje) oraz znaki interpunkcyjne z wyłączeniem niektórych użyć dywizu oraz apostrofu. Znaki interpunkcyjne będące separatorami słów uznano za osobne segmenty.

W niektórych przypadkach zdecydowano jednak za segmenty uważać ciągi krótsze od słów. Jako odrębne segmenty traktowano:

1. tzw. formy aglutynacyjne leksemu BYĆ, czyli ruchome końcówki *-(e)m*, *-(e)ś*, *-(e)śmy*, *-(e)ście*; następujące słowa reprezentują zatem po dwa segmenty (oddzielone znakiem |): *łgał* | *eś*, *długo* | *śmy*, *tak* | *em*;
2. partykuły *by*, *-ż(e)* i *-li*; następujące słowa reprezentują zatem po kilka segmentów: *przyszedł* | *by*, *napisała* | *by* | *m*, *chodź* | *że*, *potrzebował* | *że* | *by* | *ś*, *niechaj* | *że* | *ż*, *znasz* | *li*;
3. poprzyimkową nieakcentowaną formę zaimka *-ń*, np.: *do* | *ń*, *ze* | *ń*;
4. niektóre słowa zawierające łącznik, a mianowicie:
   a) słowa typu *polsko* | - | *niemiecki*,
   b) podwójne nazwiska, np. *Kowalska* | - | *Nowakowska*,
   c) podwójne imiona, np. *Jean* | - | *Pierre*.

Natomiast za jeden segment przyjmowano:

1. formy z użyciem apostrofu, takie jak *Chomsky'ego* i *(de) l'Hospitala*;
2. skrótowce zawierające łącznik sygnalizujący odmianę, np. *PRL-u*;
3. całostki (w których dywiz pojawia się w sposób ustabilizowany) typu *ping-pong*, *t-shirt*, *kogel-mogel*;
4. nazwy typu *Kędzierzyn-Koźle*, *Kudowa-Zdrój* czy *Bielsko-Biała* (natomiast były dzielone *Kraków* | - | *Płaszów*, *Wrocław* | - | *Nowy | Dwór*, *Katowice* | - | *Bogucice*, *Rybnik* | - | *Popielów* itp.).

W niniejszym projekcie kontynuowano więc zasadę (por. Woliński i Przepiórkowski 2001: 4 oraz Przepiórkowski 2004: 19 – tam też bardziej szczegółowe uzasadnienie) przyjmowania spacji za kategoryczny wyznacznik segmentacji, tzn. słowo nie mogło zawierać segmentów oddzielonych spacją. Z jednej strony wydaje się, że rozbijanie pewnych ciągów na mniejsze jednostki znaczące nie ma sensu, bo znaczą one tylko jako całość, a z drugiej doświadczenie anotacyjne uczy, że większość takich segmentów da się (w miarę sensownie mimo wszystko) oznaczyć odpowiednim fleksemem i charakterystyką morfoskładniową. Uczciwie

przyznajmy, że dla pewnej liczby segmentów decyzje takie musiały być jednak zupełnie arbitralne, np.:

1. Czym jest *mało* w *mało kto*?
2. Jaki wymagany przypadek przypisać przyimkom w konstrukcjach typu *po polsku*, *z rzadka*?
3. Czym jest *mimo* w *mimo że*, *podczas* w *podczas gdy* lub *mniej* w *mniej więcej*?
4. Jaki rodzaj nadać *innymi* w wyrażeniu *między innymi*?

Dodajmy jeszcze, że choć spacja była jednoznacznym wykładnikiem samej segmentacji, to niektóre segmenty oznaczano, biorąc pod uwagę większe (nieciągłe) jednostki, zob. p. 6.5.8 oraz p. 6.5.13.

6.3. Tagset

6.3.1. Klasy fleksemów

Pojęcie *fleksemu* wprowadził Janusz Bień (1991). Fleksem to dla nas zbiór form jednolicie lub niemal jednolicie zróżnicowanych ze względu na właściwe im kategorie gramatyczne. Na przykład klasa leksemów o nazwie czasownik składa się między innymi z fleksemów: forma nieprzeszła, pseudoimiesłów, aglutynant czasownika BYĆ, rozkaźnik, bezosobnik, bezokolicznik, odsłownik, imiesłów przysłówkowy współczesny.

Zwróćmy jeszcze uwagę na sformułowanie *niemal jednolicie zróżnicowanych*. Pełna regularność fleksyjna fleksemów jest bowiem ideałem, do którego jednak ze względów praktycznych dążyć nie warto, prowadziłoby to bowiem do wyróżniania zbyt szczegółowych klas. Tak np. istnieją rzeczowniki nie mające form liczby pojedynczej, jak DRZWI, NOŻYCZKI, tzw. *plurale tantum*. Gdybyśmy zatem chcieli się ściśle trzymać kryterium jednorodności fleksyjnej, powinniśmy dokonać bardziej szczegółowego podziału klasy rzeczowników: na odmienne przez liczbę i na te o wartości liczby ustalonej słownikowo. Podobnie należałoby wyróżnić wśród liczebników liczebniki niewłaściwe typu *dużo*, odmienne defektywnie przez przypadek. Z drugiej strony kryteria czysto fleksyjne nie pozwalają na rozbicie nieodmiennych części mowy na odpowiednie klasy. Z tego powodu kategorie fleksyjne przyjęte w niniejszym projekcie należy uważać za pewien – mamy nadzieję, że rozsądny i użyteczny – kompromis między klasami ściśle opartymi na kryteriach fleksyjnych a tradycyjnymi częściami mowy.

Bardziej szczegółowe rozumowanie dotyczące wyodrębniania fleksemów i ich stosunku do leksemów można znaleźć w pracy Przepiórkowski 2004: 25–29. W tab. 6.1 pokazano wyodrębnione fleksemy wraz z ich przypisaniem do leksemów (por. Woliński 2003).

Tabela 6.1. Klasy fleksyjne wyodrębnione w NKJP

Leksem	Fleksem	Symbol	Przykład
Rzeczownik	rzeczownik	subst	*kot, profesorowie*
	forma deprecjatywna	depr	*profesory*
Liczebnik	liczebnik główny	num	*sześć, dużo*
	liczebnik zbiorowy	numcol	*sześcioro, trojga*
Przymiotnik	przymiotnik	adj	*polski*
	przymiotnik przyprzymiotnikowy	adja	*polsko*
	przymiotnik poprzyimkowy	adjp	*polsku*
	przymiotnik predykatywny	adjc	*wesół*
Przysłówek		adv	*bardziej, kiedy*
Zaimek	nietrzecioosobowy	ppron12	*ja, tobie*
	trzecioosobowy	ppron3	*on, jemu*
	SIEBIE	siebie	*sobą*
Czasownik	forma nieprzeszła	fin	*jadam*
	forma przyszła czasownika BYĆ	bedzie	*będę*
	aglutynant czasownika BYĆ	aglt	*-śmy*
	pseudoimiesłów	praet	*jadał*
	rozkaźnik	impt	*jadaj*
	bezosobnik	imps	*jadano*
	bezokolicznik	inf	*jadać*
	imiesłów przysłówkowy współczesny	pcon	*jadając*
	imiesłów przysłówkowy uprzedni	pant	*zjadłszy*
	odsłownik	ger	*jadanie*
	imiesłów przymiotnikowy czynny	pact	*jadający*
	imiesłów przymiotnikowy bierny	ppas	*jadany*
Czasownik typu WINIEN (forma teraźniejsza)		winien	*winna, powinni*
Predykatyw		pred	*trzeba, słychać*
Przyimek		prep	*pod, we*
spójnik	współrzędny	conj	*oraz, lub*
	podrzędny	comp	*że, aby*
Wykrzyknik		interj	*ach, psiakrew*
Burkinostka		burk	*omacku, trochu*
Kublik		qub	*nie, -ż, również*
Skrót		brev	*dr, np*
Ciało obce		xxx	*errare, humanum*
Interpunkcja		interp	*;, ., (,]*

6.3.2. Kategorie gramatyczne i ich wartości

W tab. 6.2 podano wszystkie wartości kategorii gramatycznych przyjęte w anotacji NKJP. Nie omawiamy ich w tym miejscu osobno, bowiem w następnym punkcie podajemy, jakim fleksemom jakie wartości się przypisuje.

Tabela 6.2. Kategorie gramatyczne wyodrębnione w NKJP

Liczba: (2 wartości)		
pojedyncza	sg	*oko*
mnoga	pl	*oczy*
Przypadek: (7 wartości)		
mianownik	nom	*woda*
dopełniacz	gen	*wody*
celownik	dat	*wodzie*
biernik	acc	*wodę*
narzędnik	inst	*wodą*
miejscownik	loc	*wodzie*
wołacz	voc	*wodo*
Rodzaj: (5 wartości)		
męski osobowy	m1	*papież, kto, wujostwo*
męski zwierzęcy	m2	*baranek, walc, babsztyl*
męski rzeczowy	m3	*stół*
żeński	f	*stuła*
nijaki	n	*dziecko, okno, co, skrzypce, spodnie*
Osoba: (3 wartości)		
pierwsza	pri	*bredzę*
druga	sec	*bredzisz*
trzecia	ter	*bredzi*
Stopień: (3 wartości)		
równy	pos	*cudny*
wyższy	com	*cudniejszy*
najwyższy	sup	*najcudniejszy*
Aspekt: (2 wartości)		
niedokonany	imperf	*iść*
dokonany	perf	*zajść*
Negacja: (2 wartości)		
niezanegowana	aff	*pisanie, czytanego*
zanegowana	neg	*niepisanie, nieczytanego*

Akomodacyjność: (2 wartości)		
uzgadniająca rządząca	congr rec	*dwaj, pięcioma* *dwóch, dwu, pięciorgiem*
Akcentowość: (2 wartości)		
akcentowana nieakcentowana	akc nakc	*jego, niego* *go, -ń*
Poprzyimkowość: (2 wartości)		
poprzyimkowa niepoprzyimkowa	praep npraep	*niego, -ń* *jego, go*
Aglutynacyjność: (2 wartości)		
nieaglutynacyjna aglutynacyjna	nagl agl	*niósł* *niosł-*
Wokaliczność: (2 wartości)		
wokaliczna niewokaliczna	wok nwok	*-em, beze* *-m, bez*
Wymaganie kropki: (2 wartości)		
z kropką bez kropki	pun npun	*dr* (niemianownikowe formy m1), *itp*, *ul* *dr* (mianownik m1 i wszystke formy f)

6.3.3. Kategorie przysługujące poszczególnym fleksemom

Zauważmy, że tagset NKJP ma charakter morfosyntaktyczny, a nie tylko morfologiczny. Można pokazać to choćby na kategoriach przypisywanych fleksemom nieodmiennym (np. wartość wymaganego przypadka dla przyimków) czy na kategoriach niebędących fleksyjnymi dla danych leksemów (rodzaj rzeczownika, aspekt czasownika, akomodacyjność liczebnika).

W tab. 6.3 tabeli użyto symbolu $\oplus$, jeżeli dla danej klasy dana kategoria jest morfologiczna (czyli fleksem odmienia się przez tę kategorię). Symbol $\odot$ oznacza, że pewna ustalona wartość kategorii przysługuje wszystkim formom danego fleksemu, być może odróżniając go od pewnych innych fleksemów tej samej klasy (np. formy fleksemu *odsłownik* mają przypisaną ustaloną wartość kategorii rodzaju, co odróżnia go od innych fleksemów z klasy *czasownika*).

Tabela 6.3. Kategorie gramatyczne przysługujące poszczególnym fleksemom

Fleksem	**Liczba**	**Przypadek**	**Rodzaj**	**Osoba**	**Stopień**	**Aspekt**	**Zanegowanie**	**Akcentowość**	**Poprzyimkowość**	**Akomodacyjność**	**Aglutynacyjność**	**Wokaliczność**	**Kropka**
	1	2	3	4	5	6	7	8	9	10	11	12	13
Rzeczownik	⊕	⊕	⊙										
Forma deprecjatywna	⊙[a]	⊕[a]	⊙[a]										
Liczebnik główny	⊙	⊕	⊕							⊕[b]			
Liczebnik zbiorowy	⊙	⊕	⊙[c]							⊕[b]			
Przymiotnik	⊕	⊕	⊕		⊕								
Przymiotnik przyprzym.													
Przymiotnik poprzyim.													
Przymiotnik predykatywny													
Przysłówek					⊕[h]								
Zaimek nietrzecioosobowy	⊙	⊕	⊕	⊙				⊕[d]					
Zaimek trzecioosobowy	⊕	⊕	⊕	⊙				⊕	⊕				
Zaimek SIEBIE		⊕											
Forma nieprzeszła	⊕			⊕		⊙							
Forma przyszła BYĆ	⊕			⊕		⊙							
Aglutynant BYĆ	⊕			⊕		⊙						⊕	
Pseudoimiesłów	⊕		⊕			⊙					⊕[e]		
Rozkaźnik	⊕			⊕		⊙							
Bezosobnik						⊙							
Bezokolicznik						⊙							
Imies. przys. współczesny						⊙							
Imies. przys. uprzedni						⊙							
Odsłownik	⊕	⊕	⊙[f]			⊙	⊕						
Imies. przym. czynny	⊕	⊕	⊕			⊙	⊕						
Imies. przym. bierny	⊕	⊕	⊕			⊙	⊕						
Winien	⊕		⊕			⊙							
Predykatyw													
Przyimek		⊙										⊕[g]	
Spójnik współrzędny													
Spójnik podrzędny													
Wykrzyknik													
Burkinostka													

	1	2	3	4	5	6	7	8	9	10	11	12	13
Kublik												$\oplus^{g}$	
Skrót													$\oplus$
Ciało obce													
Interpunkcja													

[a] Formy deprecjatywne mają wartość kategorii rodzaju zawsze równą m2, liczby – pl, przypadka zaś – nom (w bardzo rzadkich wypadkach także acc).
[b] Kategoria akomodacyjności przysługuje wszystkim formom wszystkich liczebników.
[c] Liczebniki zbiorowe mają rodzaj n lub m1.
[d] Kategoria akcentowości przysługuje tylko niektórym formom, np. *mi* i *mnie*.
[e] Kategoria aglutynacyjności przysługuje wyłącznie zróżnicowanym formom pseudomiesłowów, tj. formom typu *niósł/niosł-*, ale nie formom typu *nieśli*, *czytał* i *czytała*.
[f] Wszystkie odsłowniki mają rodzaj n.
[g] Tylko niektóre przyimki (*z/ze*, *nad/nade* itp.) i nieliczne kubliki (*-ż/-że*) „odmieniają się" przez wokaliczność.
[h] Kategoria stopnia przysługuje wszystkim przysłówkom odprzymiotnikowym i nielicznym innym stopniowalnym (np. *bardzo*, *bardziej*, *najbardziej*), nie zaś przysłówkom typu *gdzieś* czy *wczoraj*.

6.3.4. Podstawowe różnice między tagsetami IPI PAN a NKJP

Jak już wspomnieliśmy, w tagsecie NKJP użyto wielu rozwiązań zastosowanych w tagsecie IPI PAN, włączając w to morfosyntaktyczne definicje klas gramatycznych inspirowane pracami Zygmunta Saloniego i jego współpracowników (zob. np. Saloni i Świdziński 2001) oraz szczegółowe „fleksemowe" podejście do rozróżniania klas gramatycznych za pracą Bienia (1991). Dokonano jednak pewnej liczby modyfikacji ze względów zarówno teoretycznych, jak i praktycznych[3].

Nowe klasy, kategorie i wartości

Wykrzykniki W tagsecie NKJP dodano nową klasę interj, jednak bardzo wąsko rozumianą (por. p. 6.5.10).

Spójniki W tagsecie IPI PAN wyróżniano spójniki jako jedną klasę. W tagsecie NKJP są one podzielone na spójniki współrzędne (conj) i na spójniki podrzędne (comp).

Przymiotniki predykatywne W tagsecie NKJP dodano nową klasę fleksemów przymiotnikowych nieodmiennych o oznaczeniu adjc. Klasa ta zawiera formy przymiotnikowe mogące się pojawiać wyłącznie w kontekstach predykatywnych, por. *ciekaw* w *Jestem ciekaw, co jeszcze się zmieniło*.

[3] Większość z nich została opisana w pracy Przepiórkowskiego (2009a).

Burkinostki Wprowadzono nową klasę gramatyczną burkinostek (burk). Nazwa pochodzi z artykułu Derwojedowej i Rudolfa (2003), choć są one rozumiane nieco inaczej niż tam. Należą do niej segmenty, które w korpusie IPI PAN były oznaczane arbitralnie jako rzeczowniki lub przymiotniki, a których dystrybucja ogranicza się do ściśle określonego segmentu (por. p. 6.5.9).

Skróty Dodano nową klasę fleksemów – skróty (brev). Ponadto wprowadzono dla nich techniczną kategorię wymagania kropki, przyjmującą dwie wartości: pun oraz npun. Wartość ta może być zmienna dla jednego skrótu, np. skrót *dr* może wymagać kropki (*Spotkałem się z dr. Kowalskim*) lub nie (*Poproszony został o to dr Kowalski*). Kolejną różnicą między tagsetami jest lematyzacja (hasłowanie) – w NKJP formą hasłową skrótu jest jego rozwinięcie.

Liczebniki zbiorowe W tagsecie NKJP przywrócono klasę numcol. Choć pojawia się ona w niektórych publikacjach dotyczących korpusu IPI PAN, to jednak faktycznie nie została w nim zastosowana.

Różnice między klasami, kategoriami i wartościami

Przysłówki a kubliki W NKJP przysłówki odprzymiotnikowe oraz stopniowalne mają zawsze przypisaną kategorię stopnia (oznaczane są przez adv:pos, adv:com lub adv:sup). Nie mają zaś przypisanej takiej kategorii tradycyjne przysłówki nieodprzymiotnikowe i niestopniowalne (oznaczane przez adv), jak *gdzie* lub *wczoraj*; te ostatnie są w korpusie IPI PAN przypisane do klasy kublików.

Ciała obce W tagsecie IPI PAN rozróżniano dwa typy ciał obcych: xxs – segmenty, które zajmują nominalną pozycję (i którym mogą być przypisane przypadek, liczba i rodzaj) oraz xxx – pozostałe obce wyrażenia. W tagsecie NKJP pozostawiono wyłącznie xxx (por. p. 6.5.13).

Formy nierozpoznane To klasa fleksemów nierozpoznanych w trakcie automatycznej analizy morfologicznej (tak w tagsecie IPI PAN). W ręcznej anotacji milionowego podkorpusu NKJP klasa taka nie występuje[4].

Pozostałe różnice Na koniec podamy dwie inne różnice między tagsetami. Po pierwsze, w NKJP kropka jest zawsze osobnym segmentem (także w skrótach, inicjałach, liczebnikach porządkowych zapisywanych liczbowo). W korpusie IPI PAN kropka była osobnym segmentem wyłącznie wtedy, gdy znajdowała się na końcu zdania. Po drugie, w NKJP stopień wyższy oznaczany jest przez com; używane do tego celu oznaczenie z IPI PAN (comp) przyjęto bowiem dla spójników podrzędnych.

[4] Natomiast jest ona obecna w automatycznej anotacji pełnego korpusu NKJP.

6.4. Lematyzacja

W tab. 6.4 podajemy, jakie formy hasłowe przypisywano odpowiednim fleksemom. Zauważmy przy tym, że zestawienie to przynosi dodatkową informację o tym, jakie fleksemy grupowane są w jeden leksem. Na przykład do leksemu przymiotnikowego w NKJP należy zarówno fleksem przymiotnikowy (adj), jak też fleksem przyprzymiotnikowy (adja), fleksem przymiotnika poprzyimkowego (adjp) czy fleksem przymiotnika predykatywnego (adjc) – wszystkim im przypisywana jest ta sama forma hasłowa.

Tabela 6.4. Lematyzacja w NKJP

Klasa	Symbol	Forma podstawowa	Przykład
Rzeczownik	subst	mianownik liczby pojedynczej (mnogiej dla *plurale tantum*)	WODA, SKRZYPCE
Forma deprecjatywna	depr	mianownik liczby pojedynczej odpowiedniego rzeczownika	PROFESOR
Liczebnik główny	num	mianownik rodzaju męskiego rzeczowego	PIĘĆ, DWA
Liczebnik zbiorowy	numcol	mianownik rodzaju męskiego rzeczowego odpowiedniego liczebnika głównego	PIĘĆ, DWA
Przymiotnik	adj	mianownik liczby pojedynczej rodzaju męskiego	POLSKI
Przymiotnik przyprzymiotnikowy	adja	mianownik liczby pojedynczej rodzaju męskiego	POLSKI
Przymiotnik predykatywny	adjc	mianownik liczby pojedynczej rodzaju męskiego	ZDROWY
Przymiotnik poprzyimkowy	adjp	mianownik liczby pojedynczej rodzaju męskiego	POLSKI
Przysłówek	adv	forma stopnia równego lub jedyna forma tego fleksemu	DOBRZE, BARDZO, WCZORAJ

Klasa	Symbol	Forma podstawowa	Przykład
Zaimek nietrzecioosobowy	ppron12	mianownik	ja, wy
Zaimek trzecioosobowy	ppron3	mianownik liczby pojedynczej	on
Zaimek siebie	siebie	siebie	siebie
Forma nieprzeszła	fin	bezokolicznik	czytać
Forma przyszła być	bedzie	być	być
Aglutynant być	aglt	być	być
Pseudoimiesłów	praet	bezokolicznik	czytać
Rozkaźnik	impt	bezokolicznik	czytać
Bezosobnik	imps	bezokolicznik	czytać
Bezokolicznik	inf	bezokolicznik	czytać
Imies. przys. współczesny	pcon	bezokolicznik	czytać
Imies. przys. uprzedni	pant	bezokolicznik	przeczytać
Odsłownik	ger	bezokolicznik	czytać
Imies. przym. czynny	pact	bezokolicznik	czytać
Imies. przym. bierny	ppas	bezokolicznik	czytać
Winien	winien	forma męska liczby pojedynczej	powinien
Predykatyw	pred	jedyna forma tego fleksemu	warto
Przyimek	prep	forma niewokaliczna lub jedyna forma tego fleksemu	na, przez, w
Spójnik współrzędny	conj	jedyna forma tego fleksemu	oraz
Spójnik podrzędny	comp	jedyna forma tego fleksemu	że, bowiem
Wykrzyknik	interj	jedyna forma tego fleksemu	psiakrew
Burkinostka	burk	jedyna forma tego fleksemu	trochu
Kublik	qub	jedyna forma tego fleksemu	nie, -że, się
Skrót	brev	forma hasłowa rozwinięcia	doktor, ulica, i tym podobne
Ciało obce	xxx	ten sam obiekt	errare
Interpunkcja	interp	ten sam obiekt	;, ., (,]

Poniżej podajemy kilka szczegółowych decyzji dotyczących lematyzacji:

1. Jako formę hasłową skrótowców przyjęto ich formy mianownikowe, a nie ich rozwinięcia[5], np.: forma *PRL-u* w ciągu *o PRL-u* otrzymywała formę hasłową PRL.
2. Czasowniki WINIEN oraz POWINIEN (wydzielone w osobny fleksem) nie mają bezokolicznika, w związku z tym jako forma podstawowa służy odpowiednio *winien* lub *powinien*.
3. Formą hasłową skrótu jest forma hasłowa jego rozwinięcia (w wypadku rozwinięcia do jednego segmentu) lub całe rozwinięcie (w wypadku skrótów odpowiadających frazom typu *na przykład*, *i tym podobne*, *i inne* itd.); w wypadku skrótów, których rozwinięcie może przyjmować różne wartości przypadka lub rodzaju gramatycznego, przyjmowano mianownik i rodzaj m1, np. *tzw* we fragmencie *poszedłem z tzw. poczwarnicą* oznaczono formą hasłową TAK ZWANY oraz tagiem brev:pun.
4. Wyjątek od powyższego stanowiły inicjały, choć znakowane przez brev – ich formą podstawową były one same, nie rozwinięcia.
5. Forma hasłowa liczby zapisanej cyframi arabskimi lub liczby rzymskiej to ta sama liczba.
6. Wielką literą rozpoczynano formy hasłowe nazw geograficznych (w tym formy hasłowe pisanych wielką literą członów ulic), oficjalnych nazw państw (np. CHIŃSKA REPUBLIKA LUDOWA), imion (także imion bogów i postaci typu PAPA SMERF) oraz nazwisk. Wielką literą przyjęto hasłować też nazwy, które przy danej charakterystyce gramatycznej nie mają odpowiednika w nazwie pospolitej, np. *Wprost*, *Nie* (nazwy tygodników), bo nie istnieją takie rzeczowniki pisane małą literą, *Górnik* (klub piłkarski), bo jest rodzaju m2 w odróżnieniu od *górnika* oznaczającego osobę, *Familiada*, *Seksmisja*, *Microsoft*; zgodnie z tym małą literą hasłowano odpowiednie segmenty na przykład w: *Prokuratura Wojewódzka*, *Platforma Obywatelska*, *Samoobrona*, *Biedronka*, *Ogniem i mieczem*, *Poradnik Językowy* oraz *Gazeta Wyborcza*.

6.5. Zasady znakowania

6.5.1. Rzeczowniki

Za rzeczowniki (subst) uznano fleksemy odmieniające się przez przypadek i – z wyjątkiem rzeczowników *plurale tantum* (np. SPODNIE, WUJOSTWO) i *singulare*

[5] Postąpiono tu więc inaczej niż przy skrótach, ale nie powinno to dziwić. Skrótowce funkcjonują na prawach leksemów, skróty są jedynie substytutem leksemu. Skrótowce się czyta tak, jak są zapisane, skróty zaś się rozwija. Ponadto większość skrótowców się odmienia.

tantum (np. CO, TO) – przez liczbę; każdy rzeczownik miał również przypisywaną wartość rodzaju. W ustalaniu rodzaju rzeczowników pomagały następujące konteksty testowe:

(6.3) *Widzę jednego ____ z tych, których lubię.* m1
(6.4) *Widzę jednego ____ z tych, które lubię.* m2
(6.5) *Widzę jeden ____.* m3
(6.6) *Widzę jedno ____.* n
(6.7) *Widzę jedną ____.* f

W wypadku rzeczowników *plurale tantum* formy łączące się z męskimi formami czasownikowymi w czasie przeszłym uznano za rzeczowniki rodzaju m1 (np. WUJOSTWO), pozostałe zaś – za rzeczowniki rodzaju n (np. SKRZYPCE, SANIE, POMYJE). Warto może jeszcze wspomnieć, że wartość rodzaju segmentów zależnych składniowo od rzeczownika przypisywana była zgodnie z wartością rodzaju rzeczownika, np. w zdaniu *Trzy małe pieski spały* wszystkie cztery formy mają więc przypisany rodzaj m2.

Za rzeczowniki uznane zostały także skrótowce, np. PAN, NRF, LO.

Osobny problem stanowiło zróżnicowanie form mianownika liczby mnogiej rzeczowników męskoosobowych na te, które łączą się z zaimkiem w formie *ci* (*profesorowie, emeryci*), i na te, które łączą się z zaimkiem w formie *te* (*profesory, emeryty*). Dla tego zróżnicowania Bień i Saloni (1982) zaproponowali wprowadzenie kategorii deprecjatywności, a dokładniejsze jej omówienie i argumentację zawarł Saloni w późniejszym artykule (Saloni 1988). W NKJP takie deprecjatywne formy rzeczowników uznano za osobny fleksem, odmienny defektywnie przez przypadek i mający ustaloną liczbę mnogą. Więcej na ten temat oraz dyskusję stanowisk na temat deprecjatywności (zob. Szałkiewicz 2010).

6.5.2. Liczebniki

Liczebniki główne (num) wydzielono jako klasę fleksemów o ustalonej liczbie (zawsze mnogiej[6]) i odmieniających się przez przypadek i rodzaj. Klasa ta nie obejmuje zatem tradycyjnych liczebników odmiennych przez liczbę, a więc porządkowych (DRUGI, TYSIĘCZNY), krotnych (DWUKROTNY, TYSIĄCKROTNY) itp. – uznajemy je za leksemy przymiotnikowe. Oprócz typowych liczebników głównych (DWA, PIĘĆ DZIEWIĘĆSET), za liczebniki główne uznano także: OBA, OBYDWA, gdyż odmieniają się jak DWA. Wszystkie te liczebniki można nazwać liczebnikami właściwymi. Za liczebniki główne, o nieco innej charakterystyce składniowej, a więc niewłaściwe,

[6] Argumenty za tym, że także formy liczebnikowych leksemów PÓŁ, PÓŁTORA i ĆWIERĆ mają mnogą wartość (selektywnej) kategorii liczby podane zostały w artykule Przepiórkowskiego (2006).

uznano fleksemy typu TYLE, ILE, WIELE (a także WIĘCEJ i NAJWIĘCEJ), KILKA, gdyż odmieniają się podobnie jak STO, oraz defektywne fleksemy typu DUŻO, MAŁO (a także MNIEJ, NAJMNIEJ) i TROCHĘ.

Liczebniki mają też kategorię gramatyczną akomodacyjności[7] określającą rodzaj związku pomiędzy daną formą liczebnikową a formą rzeczownikową. Na podstawie wartości tej kategorii można podzielić formy wyrazowe fleksemów liczebnikowych na dwie klasy akomodacyjne. Do pierwszej należą formy, które łączą się z rzeczownikami o innej wartości przypadka – formy rządzące rec; do drugiej zaliczymy te, które łączą się z formami nominalnymi o tej samej wartości przypadka – formy uzgadniające congr.

W osobną klasę wyodrębnione zostały liczebniki zbiorowe (numcol) – są one odrębnymi fleksemami leksemów liczebnikowych. Fleksemy liczebnikowe zbiorowe mają stałą wartość kategorii liczby, zawsze równą pl, oraz wartość kategorii rodzaju równą n lub m1, w zależności od rodzaju rzeczownika, np. *pięcioro dzieci*, *pięcioro skrzypiec* (oba n), *dwoje wujostwa* i *sześcioro studentów* (oba m1). Podobnie jak liczebniki główne, liczebniki zbiorowe mają kategorię akomodacyjności, lecz rozkład form rządzących i uzgadniających jest w tym wypadku trochę inny.

6.5.3. Przymiotniki

Wyróżniono 4 klasy fleksemów przymiotnikowych (adj, adja, adjc i adjp). Przymiotniki (adj) odmieniają się przez liczbę, przypadek, rodzaj i (nie wszystkie) przez stopień. Stopień przymiotników niestopniowalnych ustalano jako równy (pos). Przymiotniki przyprzymiotnikowe (adja) i przymiotniki predykatywne (adjc) to klasy fleksemów nieodmiennych i niesamodzielnych, występujące wyłącznie w specyficznych konstrukcjach. Przymiotniki przyprzymiotnikowe to fleksemy zawierające formy takie jak *polsko* w *polsko-niemiecki*. Przymiotniki predykatywne to niewielka, zamknięta klasa fleksemów typu ZDRÓW, CIEKAW czy GODZIEN; ich cechą jest występowanie wyłącznie w konstrukcjach predykatywnych (np. *jestem ciekaw*, ale nie **ciekaw człowiek*). Popatrzmy na przykłady oznaczeń:

(6.8) *Przyszedł do mnie pewien pan.* – PEWIEN adj:sg:nom:m1:pos (=*jakiś*).
(6.9) *Nie jestem pewna tego.* – PEWNY adj:sg:nom:f:pos.
(6.10) *Jak możesz być pewien Marka?* – PEWNY adjc.
(6.11) *Jak możesz być pewny Marka?* – PEWNY adj:sg:nom:m1:pos.

[7] Kategorię tę wprowadzili Bień i Saloni (1982). Została ona szczegółowo omówiona w pracach Wolińskiego (2003) i Przepiórkowskiego (2004).

Przymiotniki poprzyimkowe (adjp) to także klasa fleksemów nieodmiennych i niesamodzielnych, występujących wyłącznie w specyficznych (acz nierzadkich) konstrukcjach, tj. we frazach przyimkowych, np. *bez mała*, *co gorsza*, *do późna*, *od niedawna*, *na prawo*, *po ludzku*, *po prostu*, *po swojemu*, *po wojskowemu*, *w lewo*, *za młodu*, *z bliska*, *z rzadka*.

6.5.4. Przysłówki

Przysłówki (adv) to klasa fleksemów odmiennych co najwyżej przez stopień i/lub pochodzących od przymiotnika. Są to więc wyrazy stopniowalne właściwie, np. *cicho*, *ciszej*, *najciszej*, i opisowo, np. *gorzko*. Do tej klasy zalicza się także wyrazy stopniowalne niekoniecznie pochodzące od przymiotnika: *bardzo*, *bardziej*, *najbardziej*. Stopień odprzymiotnikowych przysłówków niestopniowalnych, podobnie jak odpowiadających im przymiotników, ustalany jest jako równy (adv:pos).

Dodatkowo do klasy tej należą inne (niestopniowalne, nieodprzymiotnikowe) leksemy tradycyjnie uznawane za przysłówki, np. GDZIE, GDZIEŚ, WCZORAJ, TAM, WRAZ itp. Takie przysłówki znakowano jako samo adv, bez bez podawania wartości stopnia.

Kryterium odróżniające przysłówki od kublików było przede wszystkim dystrybucyjne. Przysłówki nie mogą bezpośrednio modyfikować rzeczowników (ale mogą gerundia) ani liczebników właściwych (ale mogą niewłaściwe, np. *bardzo* w *bardzo dużo wyjątków*), natomiast zwykle modyfikują formy czasownikowe i odczasownikowe, przymiotniki i przysłówki.

6.5.5. Zaimki

Do klasy zaimków nietrzecioosobowych (ppron12) zaliczono cztery fleksemy: JA, MY, TY, WY. Każdy element tej klasy ma ustaloną liczbę (sg dla JA i TY, pl dla MY i WY) oraz osobę (pri dla JA i MY, sec dla TY i WY), odmienia się natomiast przez przypadek, rodzaj (bez zróżnicowania form) oraz akcentowość (w wypadku tych form, dla których opozycja akcentowości istnieje).

Zaimek trzecioosobowy (ppron3) to jednoelementowa klasa fleksyjna zawierająca fleksem o formie podstawowej *on*. Fleksem ten ma ustaloną osobę (ter) i jest odmienny przez przypadek, rodzaj i liczbę oraz przez kategorię akcentowości i poprzyimkowości, na przykład:

jego	ON	ppron3:sg:gen.acc:m1.m2.m3:ter:akc:npraep
go	ON	ppron3:sg:gen.acc:m1.m2.m3:ter:nakc:npraep
niego	ON	ppron3:sg:gen.acc:m1.m2.m3:ter:akc:praep
-ń	ON	ppron3:sg:gen.acc:m1.m2.m3:ter:nakc:praep

Rodzaj zaimka ppron12 i ppron3 znakowano jak rodzaj rzeczownika, do którego się odnosi. Na przykład w *Mam dość tego kota i jego głupoty* segment *jego* oznaczano jako m2 (nie: f!). Zaimki te różnią się zatem od tradycyjnie rozumianych pierwszo- i drugoosobowych form dzierżawczych, tutaj traktowanych jako formy przymiotnikowe, np. *moje* w *moje palto*: takie formy dzierżawcze uzgadniają swój rodzaj z *modyfikowanym* rzeczownikiem. Zatem tradycyjne zaimki dzierżawcze rozpadają się na przymiotniki (*mój*, *wasz* i inne zaimki dzierżawcze w 1. lub 2. os.) i zaimki osobowe w dopełniaczu (3. os.: *jego*, *jej*, *ich*).

Jednoelementowa jest klasa siebie – jej jedynym składnikiem jest fleksem SIEBIE, odmienny wyłącznie przez przypadek, zawierający trzy formy o następujących możliwych wartościach tej kategorii gramatycznej:

1. *siebie*: siebie:acc.gen
2. *sobie*: siebie:dat.loc
3. *sobą*: siebie:inst.

6.5.6. Czasowniki

Klasy fleksemów odmiennych przez liczbę i osobę obejmują klasę:

1. finitywnych form nieprzeszłych – fin; np. *zjemy* o znaczniku zjeść:fin:pl:pri:perf, *mówię* o znaczniku mówić:fin:sg:pri:imperf, *jest* o znaczniku być:fin:sg:ter:imperf;
2. form przyszłych BYĆ – bedzie; np. *będą* o znaczniku być:bedzie:pl:ter:imperf;
3. aglutynantów – aglt; np. *m* o znaczniku być:aglt:sg:pri:imperf:nwok, *ście* o znaczniku być:aglt:pl:sec:imperf:nwok;
4. rozkaźników – impt; np. *zabijmy* o znaczniku zabić:impt:pl:pri:perf, *twórz* o znaczniku tworzyć:impt:sg:sec:imperf.

Rozkaźniki mają paradygmat ograniczony do form 2. osoby liczby pojedynczej i 1. i 2. osoby liczby mnogiej. Aglutynanty (aglt) to formy *-(e)m*, *-(e)ś*, *-(e)śmy*, *-(e)ście*. Obecność litery *e* w wymienionych segmentach wyróżnia formy wokaliczne aglutynantu. Wartością ich kategorii aspektu jest zawsze imperf. Zauważmy, że mogą być dołączane nie tylko do pseudoimiesłowów (czy elementów klasy winien), lecz także do form z innych klas, np. *Głupiś*, *Abyśmy wszyscy zdrowi byli*, *Pięknieś go podsumował*. Wyrazy takie nie mają cechy aglutynacyjności i są oznaczane oddzielnie (jako przymiotniki itd.) od następujących po nich aglutynantów. Tradycyjnie rozumiane formy czasu przeszłego i trybu przypuszczającego podzielone są na formę pseudoimiesłowu lub klasy winien, kublika BY i aglutynacyjną formę czasownika BYĆ. Na przykład słowo *kupiłbym* to: *kupił* kupić:praet:sg:m1:perf, *by* by:qub, *m* być:aglt:sg:pri:imperf:nwok, zob. też p. 6.2.2.

Odsłowniki (ger) są formami czasownikowymi odmiennymi przez przypadek, ale nie przez liczbę (jeżeli dana forma występuje w liczbie mnogiej, to zawsze uznawana jest za rzeczownik). Problem odróżnienia odsłowników od rzeczowników nie został satysfakcjonująco rozwiązany, przyjęto jednak pewne kryteria przemawiające za interpretacją odsłownikową:

1. współwystępowanie argumentów „odziedziczonych" po czasowniku, np.: *pomaganie mu*, *więzienie go*, czy *twierdzenie, że…*;
2. współwystępowanie typowych okoliczników czasu, np. *bieganie (przez) 2 godziny*;
3. modyfikacja przysłówkiem, np. *robienie czegoś szybko*;
4. obecność agensa czynności we frazie typu *przez*, np. *przejęcie władzy przez rebeliantów*.

Ponieważ odsłowniki i imiesłowy (przysłówkowe współczesne i uprzednie oraz przymiotnikowe czynne i bierne) uznajemy za formy czasownikowe, to przysługuje im kategoria aspektu, formą podstawową zaś jest dla nich bezokolicznik. Formy imiesłowów współczesnych i czynnych (pcon i pact) cechują się aspektem niedokonanym (np. *pisząc*, *piszący*), a imiesłowów uprzednich (pant) – aspektem dokonanym (np. *napisawszy*). Imiesłowy bierne i odsłowniki mogą mieć dowolną wartość kategorii aspektu (*pisany/napisany*, *pisanie/napisanie*).

Kolejnymi wyróżnianymi fleksemami są bezosobnik (imps), np. *pisano*, *pito* i beozkolicznik (inf), np. *pisać*, *pić*. Są to nieodmienne czasownikowe klasy fleksyjne – przysługuje im wyłącznie słownikowa kategoria aspektu.

Pseudoimiesłów (praet), np. *pisał*, *piła*, wyróżnia się odmiennością przez liczbę i rodzaj, ale nie przez osobę. Niektóre formy niektórych pseudoimiesłowów mają także morfologiczną kategorię aglutynacyjności, np. NIEŚĆ:

niosł- agl
niósł nagl

Kategorii aglutynacyjności nie mają natomiast formy niezróżnicowane ze względu na możliwość dołączenia aglutynantu, np. *niosła* czy *czytał*.

Czasowniki POWINIEN i WINIEN odmieniają się nietypowo, dlatego fleksemy te należą do osobnej klasy winien. Elementy tych fleksemów spełniają funkcję form czasu teraźniejszego, choć mają budowę i cechy podobne do pseudoimiesłowów (nie uwidacznia się natomiast zróżnicowanie ze względu na aglutynacyjność). Formy osobowe tworzone są przez dołączanie aglutynantów. Formy czasu przeszłego i trybu warunkowego są analityczne.

Ponadto wprowadzono odrębną klasę predykatywów (pred). Klasa ta składa się m.in. z fleksemów BRAK, TRZEBA, CZAS, WARTO, SŁYCHAĆ, WIDAĆ, STRACH itp., które odmieniają się wyłącznie analitycznie (np. *było warto*, *warto*, *warto by*, *będzie warto*). Do predykatywów zaliczono także jeden z wyrazów TO.

6.5.7. Przyimki

Za przyimki (prep) uznano fleksemy nieodmienne mające funkcję łączącą oraz wymaganie określonego przypadka, zasygnalizowane w znaczniku przyimka, np. prep:dat dla formy *ku* w wyrażeniu *ku niemu*. Podstawowym testem na przyimkowość fleksemów nieodmiennych jest łączliwość z formami poprzyimkowymi (praep) zaimków: jeżeli formy danego fleksemu łączą się z poprzyimkowymi formami zaimków, to jest to fleksem przyimkowy, np.: *naprzeciwko nich*, a nie **naprzeciwko ich*.

Niektóre przyimki, np. w, z i przed, przyjmują wartości „kategorii gramatycznej" wokaliczności, np. znacznikiem *we* w wyrażeniu *we wtorek* jest prep:acc:wok.

Przyimek łączy się z rzeczownikiem, z formami przymiotnikowymi (*Wyglądał na zmęczonego*) czy liczebnikowymi (*Przyszli we trzech*). Uznano również, że przyimek może łączyć się z segmentami anotowanymi jako przysłówki, np. *na pewno*, *co najmniej*, *na trzeźwo*, *na zawsze*, czy *Pomalował dom na brązowo* i *Ugotował jajko na twardo* (Bańko 2001: 292).

6.5.8. Spójniki

Jako spójniki współrzędne (conj) wyróżniono fleksemy nieodmienne, mające funkcję łączącą, niewymagające określonego przypadka, łączące elementy o równorzędnych funkcjach w zdaniu, np. i, lecz, oraz, ani, więc. To samo oznaczenie przyjęto również dla składników spójników nieciągłych, w których każdy segment tagowano jako conj. Oto dwa przykłady takich spójników nieciągłych:

1. *nie tylko…, lecz/ale także/też/również/i…*;
2. *(zarówno)…, jak i/też/również…*

To dość wyjątkowe postępowanie skłania do zastanowienia, czy nie warto wprowadzić osobnego tagu dla oznaczania składników spójników złożonych. Nasuwa się też myśl ogólniejsza: może należałoby zastanowić się również nad innymi oznaczeniami dotyczącymi segmentów składających się na większe jednostki języka. Przy obecnej anotacji morfoskładniowej unikano jednak takich rozwiązań, brano bowiem pod uwagę następujące później znakowanie składniowe, gdzie pewne ciągi zostały połączone (np. ciągi *o tyle* i *o ile* w jeden spójnik) na poziomie słów syntaktycznych.

Poza tym wyróżniono klasę spójników podrzędnych (comp) – jednostek nieodmiennych, mających funkcję łączącą, niewymagających określonego przypadka, wprowadzających zdanie podrzędne, np. że, aby, bowiem.

6.5.9. Burkinostki

Burkinostki (burk) to jednostki nieodmienne, których dystrybucja jest ograniczona do ścisłego sąsiedztwa innego określonego segmentu, np. *trochu* w *po trochu* (ale już nie *prostu* czy *polsku* w *po prostu* czy *po polsku* – skoro istnieją przymiotniki PROSTY i POLSKI, to są to formy adjp), *naprzeciwka* w wyrażeniu *z naprzeciwka*, *oścież* w *na oścież* itp., ale także – w drodze wyjątku – *zamian* w wyrażeniu *w zamian*, choć istnieje także rzeczownikowy segment *zamian* (mający jednak inny przypadek niż przypadki wymagane przez przyimek *w*). Jest to stosunkowo niewielka klasa leksemów; nie należą do niej części nazw własnych. A zatem oba segmenty *Burkina* i *Faso* w zdaniu *Zobaczyłem na własne oczy Burkina Faso* oznaczano jako subst:sg:acc:n (z formami hasłowymi BURKINA i FASO). Większość burkinostek to człony wyrażeń z segmentem izolowanym, który kiedyś był pełnoprawnym rzeczownikiem (np. *po omacku*, *z oddali*, *na pohybel*), choć nie musi to być regułą (por. *z kretesem*; choć *kretes* w polszczyźnie nie istniał).

6.5.10. Wykrzykniki

Wykrzykniki (interj) rozumiano bardzo wąsko z tego podstawowego powodu, że w zasadzie każde słowo może zostać użyte jako pragmatyczny wykrzyknik. Za wykrzykniki uznano więc:

1. te segmenty, które mogą być użyte wyłącznie w takiej funkcji, na przykład *ach*, *och*, *oj*, *psiakrew*;
2. wyjątkowo formy, których inne interpretacje nie są związane z ich użyciami wykrzyknikowymi, np. *a* (może być także spójnikiem i skrótem), ale nie *tak*, *stop* czy *cholera*;
3. segmenty onomatopeiczne typu *muu* i *kukuryku*.

6.5.11. Kubliki

Kubliki (qub) to niespójna semantycznie i dystrybucyjnie klasa leksemów nieodmiennych (z wyjątkiem partykuł *-ż/-że*, *z/ze* „odmiennych" przez wokaliczność). Są nimi głównie te segmenty, z których każdy może modyfikować różne klasy, w tym rzeczowniki. A więc kublikiem jest *nawet*, gdyż poprawne są ciągi *nawet on*, *nawet po pijaku*, *nawet wczoraj*, *nawet poprosił*, *nawet biały* itp., ale nie segment *bardzo*, m.in. dlatego, że nie modyfikuje rzeczowników. To właśnie kryterium modyfikacji rzeczownika pozwalało odróżniać kubliki od przysłówków również tam, gdzie morfologiczna postać słowa zdawałaby się wskazywać na przysłówek. Dlatego *głównie* przyjęto zawsze oznaczać jako kublik, por. *Lubi czytać książki,*

głównie te z górnej półki / po niemiecku / dobre / kryminały. Konsekwentnie też odróżniano odpowiednie użycia słowa *prawdopodobnie*. Możemy mieć tu przysłówek stopniowalny (=*wiarygodnie*), por. *Zabrzmiało to dość prawdopodobnie. Zabrzmiało to prawdopodobniej, niż wczoraj. Zabrzmiało to najprawdopodobniej ze wszystkich twoich wypowiedzi.*, ale częściej wystąpi kublik (=*przypuszczalnie, zapewne*), por. *Zrobił to prawdopodobnie Marek. Spotkamy się prawdopodobnie wiosną*.

Jako kubliki oznaczano też m.in. następujące słowa:

1. *się* (każde wystąpienie);
2. *niech* (każde wystąpienie);
3. *nie* (jako negacja);
4. partykuły *-li*, *-ż/-że*, *-by*;
5. modyfikatory liczebności typu *z* (*z 5 kilo*), *około* (*około tysiąc osób przyszło*), *koło* (*koło tuzina*), *gdzieś* (*gdzieś ze dwieście od razu wyszło*), *bodaj*, *aż*, *dokładnie*, *ponad* itp.;
6. segmenty *za*, *zbyt*, *dość*, *dosyć* itp. w kontekstach typu *za/zbyt/dość/dosyć dużo/często/wysoki*;
7. wszystkie użycia segmentów *już* (*Studia już skończylem*, *4 tysiące akapitów to już jest coś*; *Już, już!*) i *jeszcze* (*To jeszcze nie koniec*, *Jeszcze czego!*).

6.5.12. Skróty

Skróty (brev) (ale nie skrótowce, por. p. 6.5.1) uznano za osobną klasę gramatyczną. Dla klasy tej przewidziano jedną specjalną kategorię wymagania kropki, o wartościach pun (kropka wymagana, np. dla segmentu *dr* w zdaniu *Widziałem się z dr. Kowalskim*) i npun (kropka niewymagana, np. w zdaniu *Dr Kowalski przybył*). Informacja o wymaganiu bądź niewymaganiu kropki jest niezależna od tego, czy faktycznie ona w tekście występuje. Na przykład w zdaniu *To są różne zwierzęta, np konie.* skrót *np* powinien być oznaczony jako brev:pun.

6.5.13. Ciała obce

Ciała obce (xxx) to segmenty pochodzenia obcego, będące częściami dłuższych cytatów czy sentencji, a więc nie wchodzące w bezpośrednie oddziaływania z segmentami polskimi w tekście. Na przykład każdy z trzech pierwszych segmentów w zdaniu *My czeczinskije bojewiki – mówią o sobie z dumą dwaj 13-latkowie.* został oznaczony jako xxx.

Gdy jednak były to pojedyncze słowa lub bardzo krótkie ciągi słów obcych, wchodzące w interakcje gramatyczne ze słowami polskimi, anotowano je jak słowa polskie, nadając wszystkim segmentom w tym krótkim ciągu ten sam znacznik, a mianowicie odpowiedni dla danej pozycji w zdaniu i elementu głównego tej

grupy obcej. Formami hasłowymi takich segmentów były te same segmenty. W razie wątpliwości co do rodzaju zdecydowano przyjmować arbitralnie m3. Na przykład w zdaniu *Kupiła Washington Post*, zarówno *Washington*, jak i *Post* oznaczono jako subst:sg:acc:m3. Takie krótsze ciągi mogły rzecz jasna należeć do różnych klas gramatycznych, por. *warunkiem sine qua non jest ustalenie* – adj, *metoda in vitro* – adj, *zapłodnić in vitro* – adv, *ocenić in plus* – adv, *różnice in plus* – adj.

Był to drugi przypadek, obok spójników złożonych, wyjścia poza jeden segment i anotacji jednym znacznikiem segmentów tworzących większą całość.

6.5.14. Przykład anotacji

Podamy tu jeszcze przykład zdania oznaczonego zgodnie z omawianym powyżej tagsetem i zasadami. Oto to zdanie: *Albowiem John Roder otrzymał propozycję objęcia nowego laboratorium przy znanym szpitalu Mount Sinai Hospital i po prostu z całym zespołem z Kingston przenosił się do Toronto, zabierając sześcioro pracowników.*

- Albowiem – albowiem comp
- John – John subst:sg:nom:m1
- Roder – Roder subst:sg:nom:m1
- otrzymał – otrzymać praet:sg:m1:perf
- propozycję – propozycja subst:sg:acc:f
- objęcia – objąć ger:sg:gen:n:perf:aff
- nowego – nowy adj:sg:gen:n:pos
- laboratorium – laboratorium subst:sg:gen:n
- przy – przy praep:loc
- znanym – znany adj:sg:loc:m3:pos
- szpitalu – szpital subst:sg:loc:m3
- Mount – Mount subst:sg:nom:m3
- Sinai – Sinai subst:sg:nom:m3
- Hospital – Hospital subst:sg:nom:m3
- i – i conj
- po – po prep:acc
- prostu – prosty adjp
- z – z prep:inst:nwok
- całym – cały adj:sg:inst:m3:pos
- zespołem – zespół subst:sg:inst:m3
- z – z prep:gen:nwok
- Kingston – Kingston subst:sg:gen:n
- przenosił – przenosić praet:sg:m1:perf
- się – się qub
- do – do prep:gen

– Toronto – Toronto subst:sg:gen:n
– , – interp
– zabierając – zabierać pcon:imperf
– sześcioro – sześć numcol:pl:acc:m1:rec
– pracowników – pracownik subst:pl:gen:m1
– . – interp

6.6. Anotatornia i znakowanie

6.6.1. Anotatornia

W niniejszym punkcie prezentujemy kwestie techniczne i organizacyjne ręcznej anotacji morfoskładniowej podkorpusu milionowego w projekcie Narodowy Korpus Języka Polskiego za pomocą systemu Anotatornia (Hajnicz i in. 2008, Przepiórkowski i Murzynowski 2011).

NKJP obejmuje anotację na różnych poziomach lingwistycznych. Najpierw odbywał się podział na zdania i segmenty oraz anotacja morfosyntaktyczna. Zasady tych etapów znakowania zostały określone w szczegółowej instrukcji. Następnie dołączona została anotacja semantyczna – ujednoznacznianie semantyczne niektórych form tekstowych. Wszystkie te zadania były wykonywane w systemie Anotatornia. Następne poziomy anotacji (tzn. poziom słów i grup składniowych oraz jednostek nazewniczych) były realizowane już z wykorzystaniem innych narzędzi.

Ręczna anotacja to jedno z najbardziej kosztownych zadań (jeśli nie najkosztowniejsze) w rozwijaniu korpusu, ważne jest zatem, by mieć do niej odpowiednie narzędzie. Takiego narzędzia wymagała właśnie ręczna anotacja milionowego podkorpusu NKJP. Musiało ono działać jako system webowy, umożliwiać pracę online (najlepiej za pomocą różnych przeglądarek[8]), zapewniać zapisywanie wyników w trakcie pracy itd. Nie opisujemy tutaj szczegółowo tych wymagań, ponieważ czytelnik może je wywnioskować z opisu samej Anotatorni w dalszych partiach tekstu.

W momencie rozpoczęcia projektu NKJP nie znaleziono odpowiedniego narzędzia, które spełniałoby wymogi, więc zdecydowano się zaadaptować do tego celu właśnie Anotatornię. Jej prototyp (Hajnicz i in. 2008) był rozwijany we wcześniejszym projekcie prowadzonym przez IPI PAN. Obecnie jest ona udostępniona na licencji GNU General Public License pod adresem `http://nlp.ipipan.waw.pl/Anotatornia/`.

[8] Anotatornia była najintensywniej testowana z przeglądarką Firefox.

Natomiast w samym projekcie udostępniono anotatorom dwie wersje Anotatorni, ćwiczebno-testową – mającą służyć użytkownikom do ćwiczeń, zwłaszcza przed przystąpieniem do właściwej anotacji (a także do testowania rzeczy nowo zaimplementowanych) – oraz produkcyjną, wykorzystywaną do właściwej pracy.

6.6.2. Organizacja pracy

W projekcie nie było podziału anotatorów na poziomy anotacji. Anotatorzy, którzy dostali do opisu dany akapit, znakowali go na wszystkich szczeblach – dzięki temu liczono na zachowanie spójności tagowania. Na każdym poziomie znakowanie było prowadzone niezależnie przez dwie osoby. Warto podkreślić, że pierwsze trzy poziomy anotacji (podział na segmenty, podział na zdania i anotacja morfosyntaktyczna) są kluczowe dla jakichkolwiek dalszych lingwistycznych etapów znakowania tekstu (choćby dla anotacji składniowej). Każdy więc z tych pierwszych trzech poziomów wykonywany był systemem „2+1" – dwóch anotatorów wprowadza odpowiednie informacje niezależnie od siebie, a ewentualne konflikty rozsądza osoba trzecia – tzw. superanotator.

System przydzielał poszczególnym anotatorom transze w sposób losowy. Każda transza składała się z 10 akapitów, tj. fragmentów tekstu zawierających zwykle kilka zdań, około 40–70 słów (rys. 6.1). 5 spośród tych akapitów otrzymywał równolegle w innej transzy inny anotator a drugie pięć akapitów trafiało do jeszcze innego anotatora. Anotatorzy nie wiedzieli, kto jest drugą osobą opracowującą dany akapit. W ten sposób unikało się współpracy przy uzgadnianiu anotacji poza Anotatornią (co było samo w sobie niedozwolone). Jeden anotator mógł pobrać jednorazowo do 25 transz.

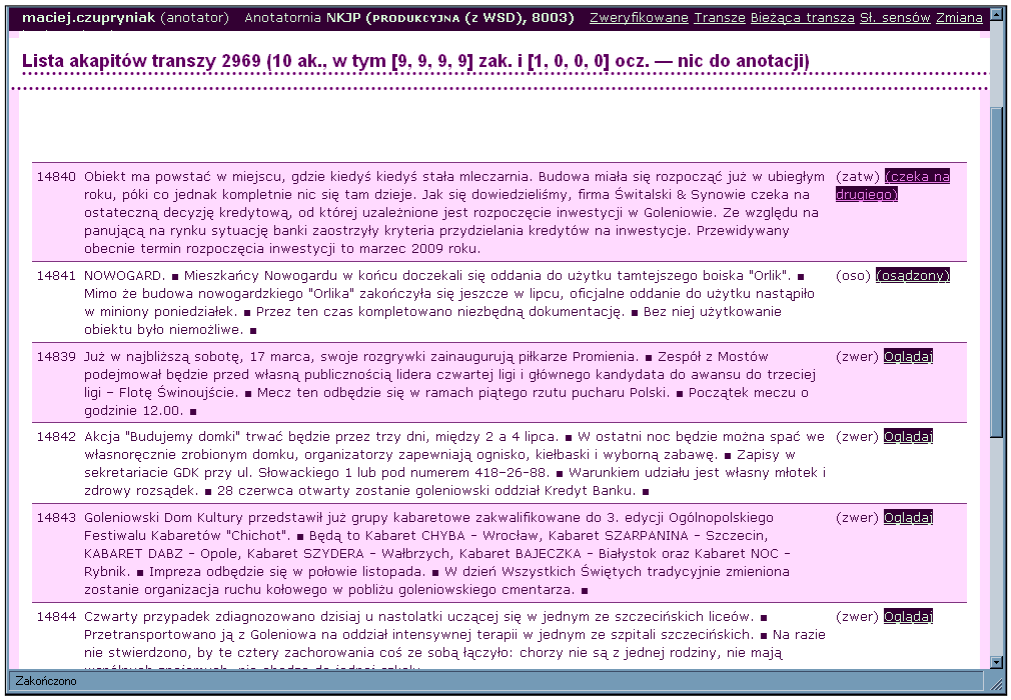
Jeśli anotatorzy zgodnie oznaczyli akapit na danym poziomie anotacji, to otrzymywał on status zweryfikowanego (zwer) i przechodził na kolejny poziom, gdzie znakowali go ci sami anotatorzy (przy czym segmentację na zdania i anotację morfoskładniową można było wykonywać równocześnie, te dwa poziomy były od siebie niezależne).

Natomiast gdy występowały rozbieżności, akapit wracał do anotatorów. Widzieli oni zaznaczony na żółto fragment, gdzie wystąpiła rozbieżność, ale wciąż nie widzieli, jak oznaczył drugi anotator (rys. 6.2).

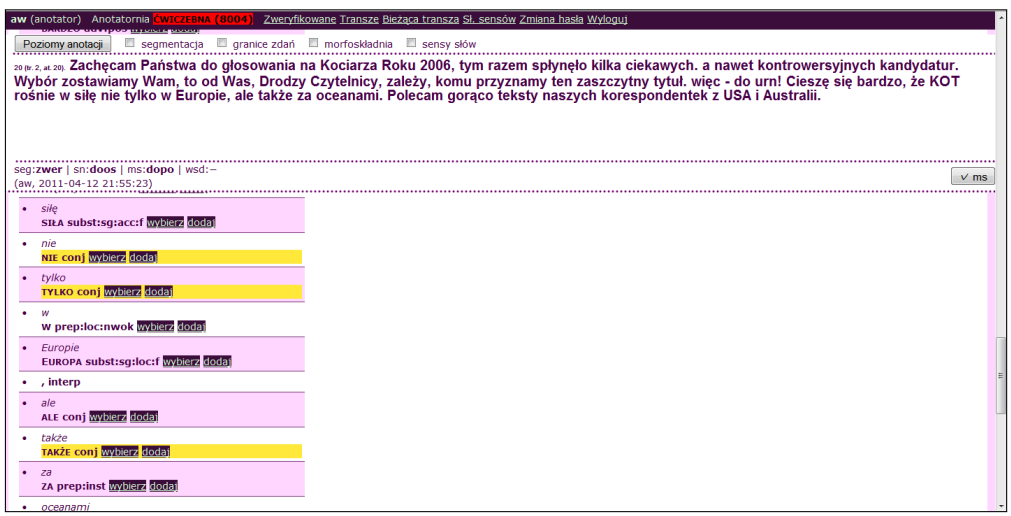
Anotatorzy mogli wtedy zmienić swój opis lub uznać, że jest on właściwy i przy nim pozostać. W każdym razie obaj musieli ponownie przemyśleć swój wybór. Jeśli po tym etapie dalej występowała niezgodność, to akapit trafiał do superanotatora, decydującego, który opis jest właściwy[9]. Dopiero wtedy anotator

[9] Mogło też zdarzyć się tak, że superanotator nadawał spornemu segmentowi jeszcze inny opis, jeśli uznał, że wybrane przez anotatorów były niewłaściwe.

Rysunek 6.1. Widok akapitów w transzy

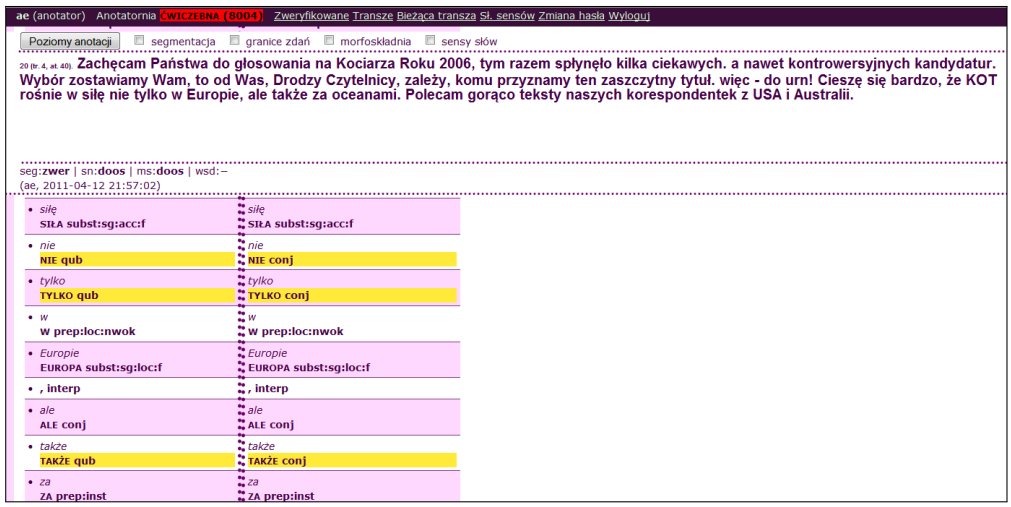


Rysunek 6.2. Niezgodności na poziomie morfoskładniowym
(akapit wraca do anotatora, który nie widzi, jak go oznaczyła druga osoba)



mógł zobaczyć, jak sporne segmenty oznaczyła druga osoba, która otrzymała ten sam akapit do anotacji (rys. 6.3). Później anotatorzy mogli również zobaczyć w zakończonym już akapicie, jaką decyzję podjął superanotator.

Rysunek 6.3. Niezgodności na poziomie morfosyntaktycznym (akapit trafił do osądu)



Jak wspomnieliśmy, anotatorzy znakowali akapity niezależnie od siebie. Stworzona została natomiast (e-mailowa) lista dyskusyjna[10], na którą można było pisać w przypadku wątpliwości czy problemów z odpowiednią anotacją. Dyskusja prowadzona na liście była widoczna dla wszystkich anotatorów.

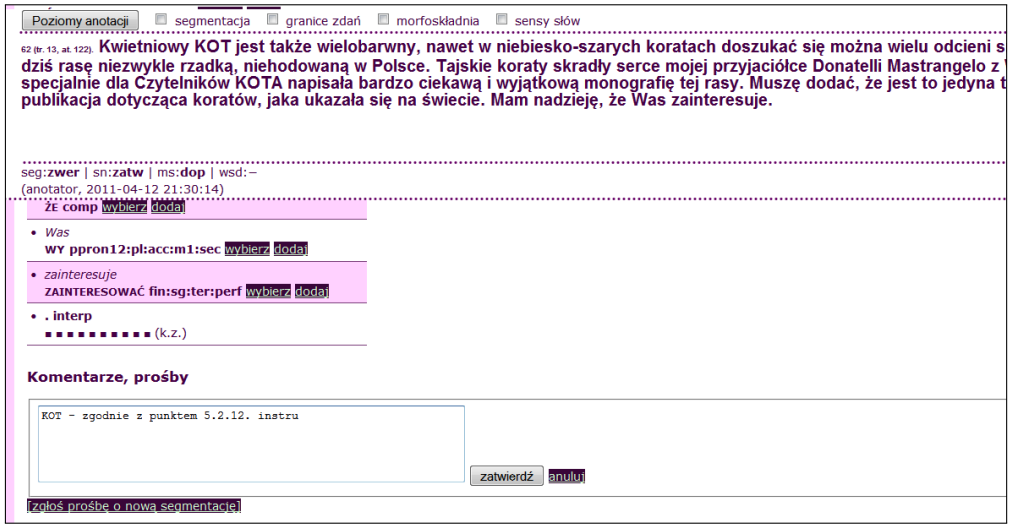
Dodatkowo pod anotowanym akapitem można było dopisać komentarz (rys. 6.4). Ich wpisywanie było obowiązkowe w wypadku ostatecznej niezgodności między anotatorami, tzn. przy trafieniu akapitu do osądu (rys. 6.3). Zadaniem anotatora było krótkie umotywowanie swojego wyboru, najlepiej poprzez odesłanie do odpowiedniego punktu instrukcji znakowania. Wpisanie komentarza w takim przypadku nie było jednak wymagane przez system. Ten obowiązek został nałożony na anotatorów w ogłoszeniu na liście dyskusyjnej.

6.6.3. Znakowanie

Jak wspomniano, za pomocą Anotatorni odbyło się znakowanie milionowego podkorpusu NKJP na obu poziomach segmentacyjnych (słowa i zdania), na poziomie morfosyntaktycznym oraz na poziomie sensów słów.

[10] A dokładniej dwie listy: jedna dla wszystkich anotatorów i jedna tylko dla superanotatorów.

Rysunek 6.4. Anotator wpisuje komentarz



Segmentacja na poziomie słów

Pierwszym zadaniem anotatora była segmentacja na poziomie słów. Jednak najczęściej dany akapit był już poprawnie mechanicznie podzielony na segmenty. W takim typowym przypadku, gdy akapit był jednoznaczny segmentacyjnie, zostawał on na poziomie segmentacji automatycznie oznaczany jako zweryfikowany. Jednak czasem anotator musiał dokonać pewnych rozstrzygnięć na tym poziomie. Dotyczyło to przede wszystkim słów z dywizem (*cza-cza* jako jeden segment, ale *biało-niebiesko-czerwony* jako pięć segmentów) oraz z aglutynantem (*gdzieś* jako jeden segment w *Gdzieś to położyłem*, zaś jako dwa segmenty w *Gdzieś to położył, co?*). Czasem zapytanie o rozstrzygnięcie segmentacji generował sam analizator morfologiczny. W tym wypadku jednak trzeba było uważać, by nie generować niepotrzebnych zapytań. Na przykład słowo *kiedyś* może składać się z dwóch form: *kiedy* i *ś*, jednak prawdopodobieństwo takiej sytuacji w tekście jest niewielkie. Gdyby system za każdym razem kazał anotatorowi dokonywać wyboru segmentacji, to skutkowałoby to dużą ilością niepotrzebnej pracy. Dlatego *kiedyś* występowało domyślnie jako jeden segment, a jeśli zachodziła potrzeba anotacji w rozbiciu na dwa segmenty, to anotator zgłaszał pod danym akapitem prośbę o segmentację. Wszystkie takie prośby musiały być zaakceptowane przez superanotatora.

Można było zgłosić prośbę o segmentację ze spacją, jak i o segmentację bez spacji. Można było również zgłosić prośbę o połączenie tego, co analizator

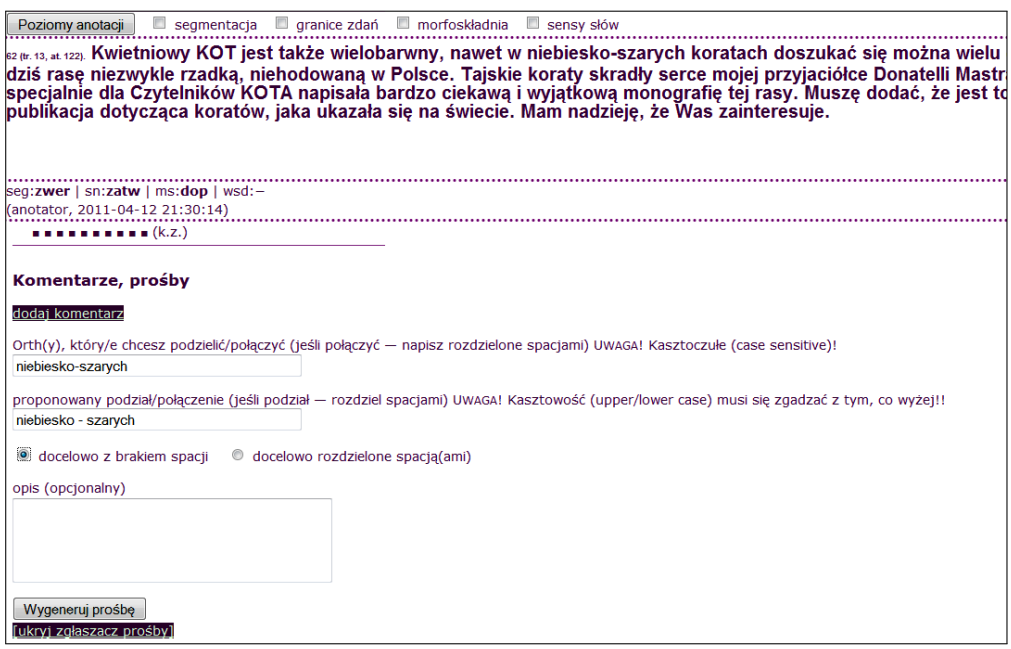
podsuwał do anotacji jako dwa segmenty (w tym z opcją likwidacji spacji między segmentami). Poniżej opiszemy dokładniej te cztery możliwości.

Segmentacja ze spacją Rozdzielanie słowa na dwa segmenty z wprowadzaniem spacji należało do rzadkości. Czyniono to tylko w wypadkach ewidentnych błędów konwersji tekstu (przede wszystkim błędów OCR, tj. złego rozpoznania tekstu w zeskanowanym dokumencie), jak np. we fragmencie *koncentracja na sobie samymi swoich doznaniach*, gdzie *i* przykleiło się do poprzedzającego słowa.

Segmentacja bez spacji Na rys. 6.5 widzimy przykład wprowadzania prośby o segmentację bez dostawiania spacji. W zdaniu *Kwietniowy KOT jest także wielobarwny, nawet w niebiesko-szarych koratach doszukać się można wielu odcieni srebra.* Anotatornia jako jeden segment uznała *niebiesko-szarych*, tymczasem zgodnie z zasadami znakowania powinny to być trzy segmenty: przymiotnik przyprzymiotnikowy (adja) *niebiesko*, dywiz, forma przymiotnikowa *szarych*. Anotator zgłasza więc prośbę o podział, ale bez spacji. W efekcie fragment ten zostanie zaanotowany jako trzy oddzielne segmenty, które jednak w tekście wciąż będą napisane łącznie.

Wprowadzenie segmentacji bez spacji pozwalało też odpowiednio znakować segmenty błędnie zapisane razem, nie rozdzielano ich w tekście zgodnie z przyjętą

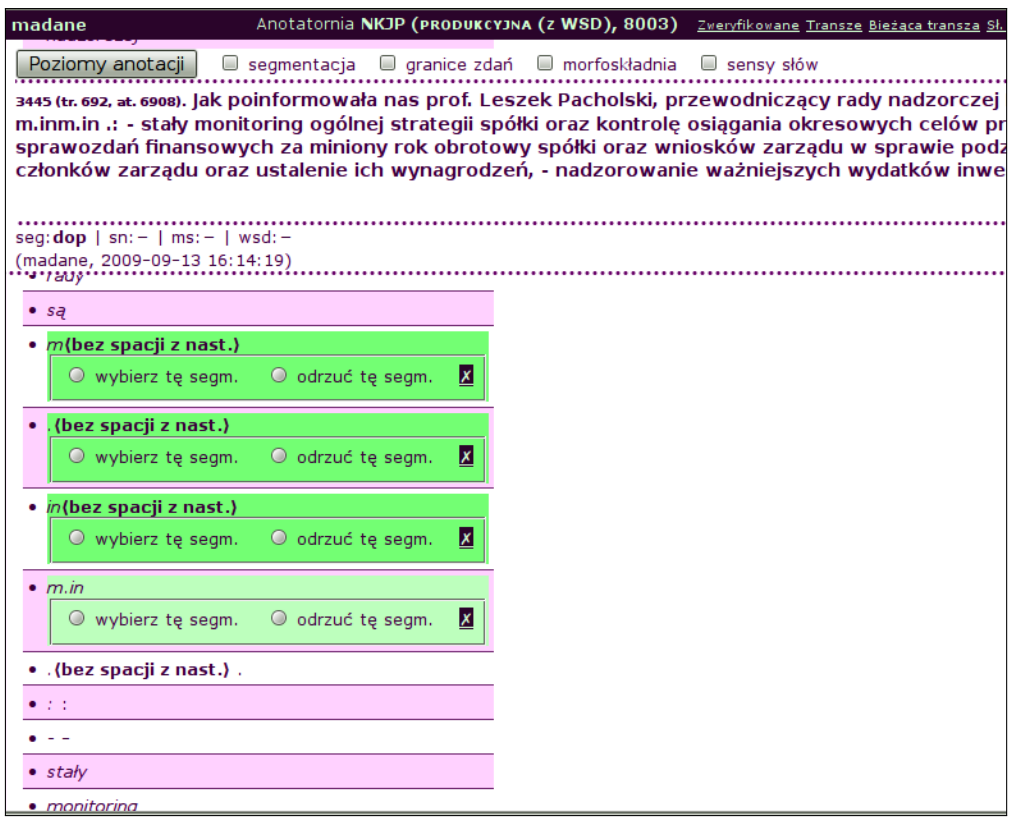
Rysunek 6.5. Prośba o segmentację



zasadą nieingerowania w teksty korpusu. W ten sposób jako dwa segmenty można było zaanotować *poprostu* we fragmencie *Osoba, która cierpi na chorobę psychiczną różni się od faceta, który poprostu rozum postradał i zabija dla przyjemności i własnej satysfakcji*. Podobnie jako dwa segmenty (rozdzielenie bez wprowadzania spacji) oznaczano liczebniki z jednostką miary (nagminnie pisane błędnie łącznie): *1km, 261ha, 30m, 600zł, 20min*.

Połączenie w jeden segment Anotator mógł zgłaszać również prośby o łączenie segmentów. Prośba taka trafiała do superanotatora. Jeśli ją zaakceptował, to akapit zostawał wysyłany do znakowania na poziomie segmentacyjnym. Przykład mamy na rys. 6.6. Anotator zgłosił do połączenia w jeden segment ciąg *m.in*, który w początkowej wersji Anotatornia uznała za trzy segmenty. Superanotator zaakceptował zgłoszenie i teraz obaj anotatorzy muszą dokonać tego samego wyboru, by ostatecznie *m.in* stało się jednym segmentem (i mogło na poziomie morfosyntaktycznym zostać oznakowane jako skrót – brev o formie hasłowej między innymi).

Rysunek 6.6. Poziom segmentacji słów



Połączenie w jeden segment z likwidacją spacji Zgodnie z zasadą nieingerowania w teksty, które trafiły do korpusu, łączenie w jeden segment segmentów rozdzielonych spacją[11] było wyjątkowe. Takie sytuacje powinny były ograniczać się do złej konwersji tekstu. Dlatego nie można było łączyć w jeden segment podkreślonych słów we fragmencie: *w prawdzie nie wiem co sie dzieje z duszą człowieka po śmierci*. Przyjęto tutaj jednak następujące, wydaje się zdroworozsądkowe, kryterium: jeśli błędy wynikają z niewiedzy użytkowników polszczyzny (np. zapisują oni słowa rozdzielnie, bo sądzą, że tak jest poprawnie), to nie ingerujemy w tekst, a jeśli błędny zapis jest raczej wynikiem nieuwagi, przypadku, błędu w konwersji tekstu, to dokonujemy ingerencji. Nie zmieniano więc zapisów typu: *w prawdzie*, *dwu osobowy*, *poza małżeński* – takie zapisy, podejrzewamy, mogą mieć pewną wartość „korpusową" (mogą nieść informację, jak piszą Polacy i jakie błędy popełniają), ale już: *po za*, *ko chany* raczej nie, więc w takich przypadkach spację generalnie likwidowano. Pojawiał się wobec tego czasem problem, w jaki sposób oznaczyć segmenty, które na poziomie tekstu są w zasadzie morfemami, a nie słowami, tak jak w zdaniu *ZSRR do roku 1980 zamierza zwiększyć liczbę studentów dwu i półkrotnie*.

Segmentacja na poziomie zdań

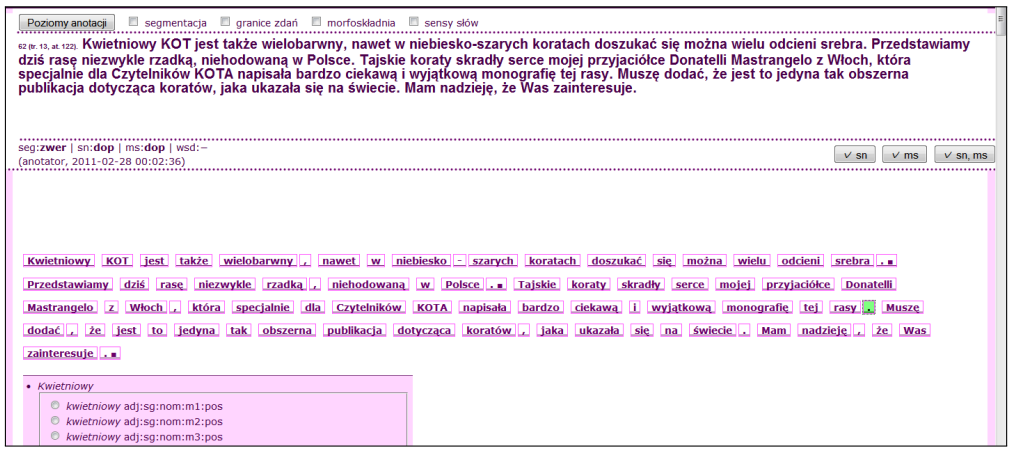
Kolejnym etapem anotacji było oznaczenie końców zdań. Anotator wprowadzał oznaczenia całkowicie samodzielnie, bez podpowiedzi ze strony systemu (jedynie na końcu akapitu był zawsze postawiony znak końca zdania). Technicznie odbywało się to w ten sposób, że każdy segment w akapicie był zamknięty w osobnej ramce i kliknięcie go umieszczało na nim znak końca zdania (kwadracik). Na rys. 6.7 widzimy, jak anotator zaznacza trzeci z kolei koniec zdania (podświetlenie na zielono). Natomiast gdy anotator próbował zakończyć tagowanie na tym poziomie, nie umieszczając znaku końca zdania na kropce, wykrzykniku lub lub innym znaku mogącym kończyć zdanie, pokazywał się komunikat *dziwię się że token(y) xxxxx nie kończą zdania* i należało jeszcze raz potwierdzić swoją decyzję.

Anotacja morfosyntaktyczna

Tekst akapitu przetwarzany był przez analizator morfologiczny Morfeusz (Woliński 2006). Interpretował on wczytane słowo jako należące do określonego leksemu lub leksemów (lematyzacja lub dehomonimizacja słaba) i przypisywał mu wszystkie możliwe zespoły parametrów morfologicznych (interpretacja morfologiczna lub desynkretyzacja słaba). Dane te były podane anotatorowi w formie wyliczenia, bez wskazania, który z opisów jest najlepszy/właściwy dla danej formy

[11] Co łączyło się oczywiście z usuwaniem tejże spacji.

Rysunek 6.7. Poziom segmentacji zdań



(kolejność propozycji nie grała żadnej roli). W ten sposób w Anotatorni wyświetlały się wszystkie interpretacje (znaczniki) możliwe do przypisania danemu segmentowi. Podkreślmy: sam analizator morfologiczny nie dokonywał tutaj żadnej dezambiguacji, oferował jedynie wszystkie możliwe interpretacje[12]. Mogło to oznaczać proponowanie samych błędnych interpretacji (nawet jeśli podawany był tylko jeden tag), jak też nieproponowanie żadnego znacznika (gdy np. nie było odpowiedniego leksemu w bazie słownej analizatora). Zadaniem anotatora było ujednoznacznienie segmentu, tj. dokonanie desynkretyzacji mocnej i dehomonimizacji mocnej, innymi słowy – rozpoznanie właściwej formy wyrazowej i właściwego leksemu. Mówiąc jeszcze prościej – anotator powinien wybrać właściwy znacznik spośród propozycji Morfeusza lub dodać własny tag, jeśli odpowiedniego wśród wyświetlonych propozycji nie było. Dla przyspieszenia wprowadzania znaczników Anotatornia podpowiadała dalszy ciąg na podstawie używanego tagsetu. Zostały też zaimplementowane pewne ograniczenia, niepozwalające na wpisanie niepoprawnych wartości, np. dla ger (gerundium) nie można było podać innego rodzaju niż n (nijaki).

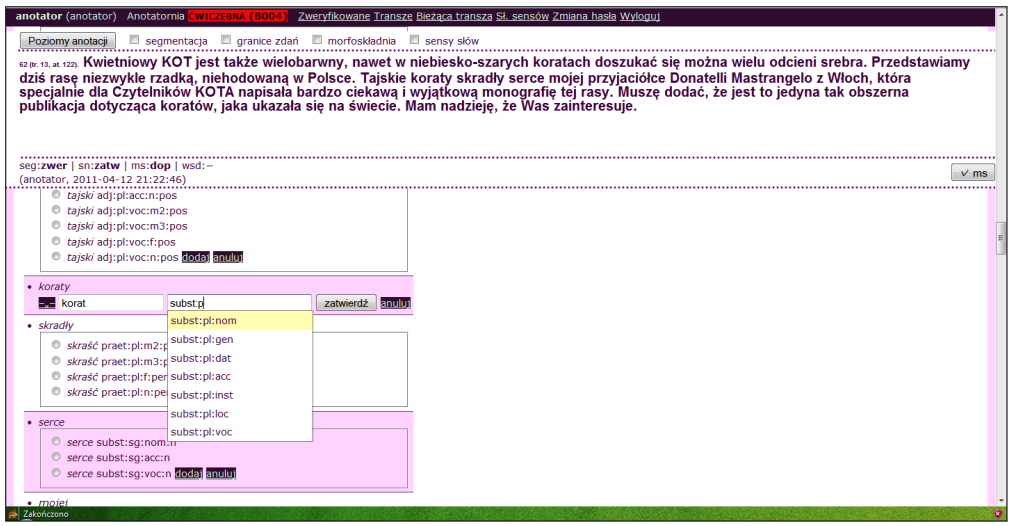
Na rys. 6.8 widzimy przykład, w którym Morfeusz nie rozpoznaje formy *koraty*. Anotator musi więc sam wpisać formę hasłową (KORAT), a następnie podać odpowiednie tagi. Po wprowadzeniu początkowych liter znacznika system proponuje dalszy wybór w postaci rozwiniętej listy.

Każdy segment musiał być ujednoznaczniony przez anotatora[13], co czasem było zadaniem niełatwym. Nie zawsze bowiem dało się ustalić jedyną poprawną

[12] Typowy analizator morfologiczny tak właśnie działa – podaje wszystkie możliwe interpretacje. Automatycznym ujednoznacznieniem zajmuje się tager części mowy.

[13] Z wyjątkiem segmentów interpunkcyjnych, które z założenia miały tylko jedną interpretację morfoskładniową: interp.

Rysunek 6.8. Anotator wprowadza nowy opis



interpretację. Nieraz trzeba było sięgać do korpusu w poszukiwaniu szerszego kontekstu, szukać poprzedniego zdania w akapicie wcześniejszym itp. Należało np. scharakteryzować nadawców, bohaterów dialogów pod względem płci, by móc poprawnie oznaczyć rodzaj choćby zaimków pojawiających się w ich wypowiedziach. Jeżeli teoretycznie kilka interpretacji mogło być w danym kontekście właściwych, należało wybrać tę, która wydawała się typowa. Na przykład, choć w zdaniu *Podał mu guzik.*, rozpatrywanym bez szerszego kontekstu, forma *podał* może mieć różne interpretacje rodzajowe (m1, m2, m3), należało wybrać formę rodzaju męskoosobowego (m1) jako najbardziej prawdopodobną. Oczywiście, jeżeli z kontekstu wynikało, że to automat podał guzik, należało wybrać formę rodzaju zgodną z rodzajem rzeczownika AUTOMAT, a więc m3.

Czasem przypisanie pewnej wartości było arbitralne, nawet jeśli znało się kontekst (tak na przykład było z rzeczownikami o obocznym rodzaju m2/m3). W tej sytuacji można byłoby wprowadzić możliwość pozostawiania kilku interpretacji. Czasem bowiem nie da się niearbitralnie wybrać tej jedynej, nawet gdy mamy pełny kontekst, pełną semantykę zdania i pełną wiedzę o świecie. Jednak wcześniejsze doświadczenia znakowania korpusu IPI PAN pokazały, że dawanie możliwości wybierania więcej niż jednej interpretacji powoduje niepożądane skutki. Anotatorzy bowiem wykorzystują tę możliwość zbyt często w najróżniejszych sytuacjach. Decyzja o wymuszaniu jednej interpretacji była więc przemyślana i poparta doświadczeniem.

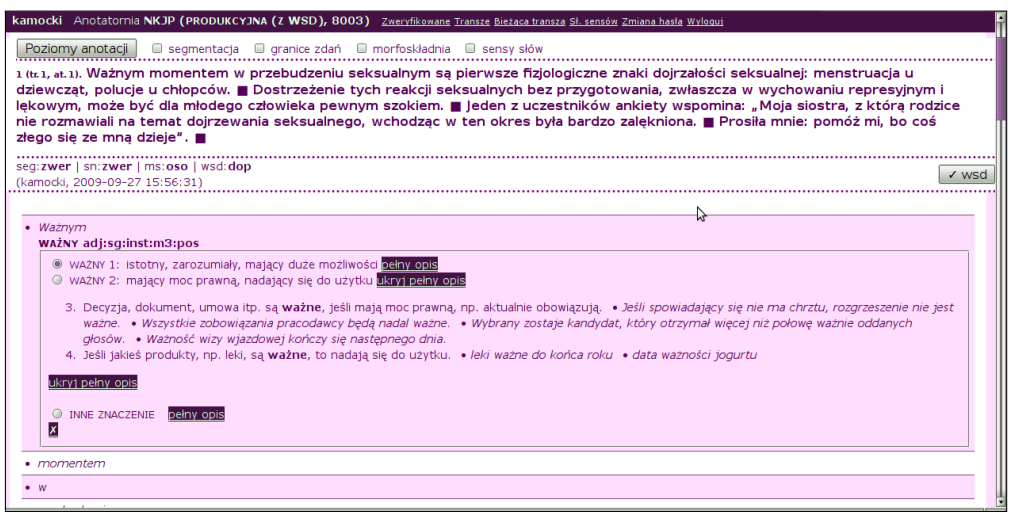
Pokażmy jeszcze ciekawy przykład, który pokazuje, jak szeroka i różnorodna wiedza była czasem potrzebna do odpowiedniego znakowania morfosyntaktycz-

nego. Otóż anotatorzy otrzymali do otagowania zdanie: *Według UEFA mistrz Polski Górnik Zabrze ma spotkać się w pierwszym meczu z Djurgardem Sztokholm 20 września w Zabrzu*. Podkreślony segment zgodnie oznaczyli jako rzeczownik rodzaju męskorzeczowego w narzędniku o formie podstawowej Djurgard. Przypadek chciał, że zdanie to trafiło do superanotatora z powodu niezgodności w innym segmencie. Superanotator dostrzegł jednak literówkę. Jako zapalony kibic piłkarski znał Djurgardens IF, szwedzki klub piłkarski z siedzibą w Sztokholmie i mógł odpowiednio poprawić anotację.

Anotacja sensów słów

Anotacja na poziomie sensów słów (wsd – *word sense disambiguation*) objęła tylko wybrane (około 100) leksemy. Odbywała się w ten sposób, że ta sama dwójka anotatorów otrzymywała do ujednoznacznienia swoje poprzednio anotowane akapity, gdy już na wcześniejszych poziomach anotacja została uzgodniona. System sam sprawdzał, czy w danym akapicie znajdowały się formy do ujednoznacznienia. Jeśli tak było, to pojawiała się przy nim zachęta [Anotuj wsd]. Oznaczanie przebiegało podobnie jak na poziomie morfosyntaktycznym, tj. anotator wybierał z rozwijanej listy sensów ten odpowiedni. Wyświetlały się one na liście jako przybliżone definicje, po kliknięciu w dany punkt można było go rozwinąć i zobaczyć szczegółowy opis wraz z przykładami (rys. 6.9). Bardzo drobne sensy wyróżniane w *Innym słowniku języka polskiego* (Bańko 2000) zostały pogrupowane w ten sposób, że podobne znaczenia znalazły się w tych samych grupach; zob. rozdz. 7. Taki ogólniejszy podział na sensy jest wskazany z punktu widzenia przetwarzania języka.

Rysunek 6.9. Anotacja na poziomie sensów słów



Znakowanie semantyczne, jako czwarty kolejny poziom, odbywało się na podstawie wcześniejszych anotacji, tzn. na poziomie wsd oznaczane były te segmenty, które wcześniej zostały przypisane do odpowiednich leksemów.

6.6.4. Z doświadczeń anotacji

O nieingerencji w teksty

Z założenia nie poprawiano tekstów oddanych do Korpusu. Pisaliśmy o tym już w punkcie dotyczącym segmentacji, zob. p. 6.6.3. Dlatego w większości przypadków należało znakować formy zastane, a nie dointerpretowane. Tak np. w zdaniu *Na miejsce tragedii zawiódł go zapach rozkładającego go ciała.*, choć możemy domyślać się, że zamiast drugiego *go* powinno być *się*, to znakowanie odbywało się zgodnie z tym, jaka forma występuje w tekście[14].

Anotatorzy mieli jednak jeszcze pewną możliwość. Dopuszczono bowiem anotację form „domyślnych" w wypadkach podejrzeń o literówkę, złą konwersję tekstu itp. Na przykład we fragmencie *kotką i tej instynktami macierzyńskimi interesowali się także biolodzy radzieccy* powinno chyba być *jej*, a nie *tej*. Wobec tego oznakowano ten segment jako on ppron3:sg:gen:f:ter:akc:npraep, uznając w tym wypadku błąd konwersji[15]. Przy podjęciu takiej decyzji obowiązkiem anotatora było umieszczenie pod akapitem komentarza o odpowiedniej, ustalonej treści. Dalszym etapem, po zakończeniu znakowania korpusu, powinno być więc odpowiednie poprawienie tekstów w tych miejscach.

Cała kwestia poprawiania czy niepoprawiania tekstów i odpowiedniej anotacji jest bardzo skomplikowana i tu ją zaledwie sygnalizujemy. Nie do końca wiadomo na przykład, co zrobić w wypadku zdania:

– *Prof. Andrzej Mackiewicz jest laureatem wielu znaczących wyróżnień: m.in. wielokrotnie Nagrody Ministra Zdrowia i Opieki Społecznej, Nagrody im. L. Hirszfejda oraz Amerykańskiego Towarzystwa Reumatologicznego.*

tzn. jaką formę hasłową nadać podkreślonemu segmentowi (skoro to błąd autora tekstu, a takich z założenia nie poprawiamy).

Zróżnicowanie tekstów a znakowanie

Należy mieć świadomość, że do ręcznej anotacji trafiały teksty różnego typu. Założenie było bowiem takie, by w korpusie anotowanym ręcznie znalazły się

[14] Weźmy też pod uwagę, że na tekstach znakowanych ręcznie uczy się automaty. Nadawanie danemu kształtowi graficznemu interpretacji niezgodnej z tym kształtem mogłoby powodować błędy w znakowaniu automatycznym.

[15] W takich wypadkach przydatna była wiedza o tym, z jakich tekstów pochodzi fragment, czy był skanowany i jakie typowe błędy zdarzają się skanerom i programom OCR.

teksty maksymalnie różnorodne (zob. rozdz. 5). Anotatorzy mieli więc do czynienia z fragmentami nie tylko z powieści, z prasy, ale i z przewodnika górskiego, z sennika, z zapisami rozmów, z tekstami z Internetu (w tym z blogów, Wikipedii itd.). Wiązały się z nimi specyficzne problemy.

Załóżmy, że chcemy mieć jakąś reprezentację tekstów z blogów czy forów internetowych. Musimy zatem liczyć się z błędami ortograficznymi (zwłaszcza w pisowni łącznej lub rozdzielnej) i rozchwianą interpunkcją. Nie dotyczy to oczywiście wszelkich tekstów internetowych tego typu – pewne z nich nie są pod tym względem gorsze od tekstów prasowych. Dziś nie jest już tak, że blogi kojarzą się z nieporadną pisaniną nastoletniej młodzieży. Piszą je uznani artyści, dziennikarze, politycy. Inna sprawa, że są to wciąż teksty nieprzechodzące z wiadomych względów przez redakcję i dlatego zdarzają się w nich błędy. Karkołomnym zadaniem byłoby jednak sprawdzanie wszystkich takich tekstów przez człowieka, by w korpusie uwzględnić wyłącznie napisane w stu procentach poprawnie. Zresztą skoro korpus ma być reprezentatywny, a dla pewnych tekstów charakterystyczny jest niższy poziom polszczyzny, to może trzeba przyjąć je takimi, jakimi są.

Teksty z Internetu to jednak nie tylko błędy pisowni. Wraz z nimi pojawiły się w anotacji np. emotikony, takie jak *:-)*, *xD* czy *:-PP*, akronimy, jak *EOT, IMHO*, nicki internetowe, jak *Szalony Odys, EkstazyGirl*, adresy stron internetowych, jak *wrzuta.pl, www.hydepark.pl* (czasem używane nawet jako odmienna nazwa, por. *Nie działa odtwarzanie na wrzucie.pl*). Należało więc zdecydować, w jaki sposób je znakować[16]. Dodatkowej decyzji wymagała także anotacja na poziomach segmentacyjnych. Emotikony łączono w jeden segment, a znak końca zdania przyjęto stawiać na emotikonie w wypadku, gdy piszący kończyli nim zdanie, nie stawiając kropki. Adresy stron internetowych również znakowano jako jeden segment.

Najwięcej problemów w anotacji pojawiło się w tekstach mówionych (tj. zapisach rozmów). Transkrypcja tych tekstów wykonana została bowiem z pewną nieuniknioną raczej dozą niespójności (por. rozdz. 4). Generalnie zapisywane były z polskimi znakami, ale zdarzały się formy typu „robie" bez ogonka. Ponadto część słów została (zgodnie z zaleceniami dla spisujących) zapisana fonetycznie[17], co rodziło problem wyboru dla nich formy hasłowej. Na przykład, gdy trafiało się na *fest lejdi* (z kontekstu wynikało, że to nazwa solarium), to nie wiadomo było, czy chodziło o *First Lady*, czy o żartobliwą nazwę *Fest Lejdi* (skądinąd całkiem ciekawą).

Zalecenia dla spisujących teksty mówione zawierały zapis o odnotowywaniu szczególnie charakterystycznej i odbiegającej od wzorcowej normy wymowy. To

[16] Emotikony przyjęto oznaczać tak, jak różnego rodzaju symbole, a więc jako rzeczowniki rodzaju nijakiego. Formą hasłową emotikonów z powtórzonym elementem czyniono emotikon „podstawowy", np. formą hasłową *;-))))* było *;-)*. Przy akronimach typu *BTW* problematyczne było rozstrzygnięcie, czy są to skrótowce, czy skróty. Ostatecznie przyjęto to drugie rozwiązanie.

[17] W tekstach tych znajdowały się więc m.in. zapisy: *alegro*, *najki*, *dolby serand*, *starłorsy*.

tłumaczy obecność *powyłanczane*, *wciągło*, *będziem*, zapewne też *dwajścia*, *dziecioki*, *łon* (*on*), *teramizu* (*tiramisu*), *ktuś* (*ktoś*), ale chyba niekoniecznie *przywieś* (*przywieź*), *pice* (*o tam pice może zjeść nie*), *o szty* (*ożty*) czy wygłosowych nosówek bez ogonków – zwykle brak nosówek na końcu czasowników (*ja przychodze, ja pracuje*). Spisujący te teksty, jak widać, mieli czasem problem z właściwym zapisem. W rezultacie otrzymaliśmy, przynajmniej w niektórych akapitach, zapisy problematyczne do interpretacji morfoskładniowej.

Kolejnym problemem w tekstach mówionych była anotacja (w tym wybór formy hasłowej) dla zapisanych ze spacją skrótów czy skrótowców. Popatrzmy na przykłady zdań:

(6.12) *a tu ci kochana w pe ka o powiedzieli że muszę konto wyczyścić do zera żeby ruszyć kredyt*;
(6.13) *telewizja polska yy nadaje kanały takie jak te fał pe sport te fał pe ha de*;
(6.14) *podglądają nas przez w w w polskie radio euro kropka pe el*[18];
(6.15) *to gramy w ce esa a później idziemy spać?*;
(6.16) *redukują emisję ce o dwa*.

Najlepszym wyjściem wydawało się w tym wypadku dopuszczenie możliwości sklejania tak zapisanych słów w jeden segment. Trzeba bowiem stwierdzić, że o sposobie ich zapisu decydowały zasady transkrypcji, więc takie ingerencje na potrzeby anotacji były wskazane. Podobnie jak połączenie w jeden segment (zlikwidowanie spacji) *na prawdę*. W tekście z Internetu taka ingerencja była niedopuszczalna i należało anotować oddzielnie dwa segmenty; było bowiem ważne, żeby taki błąd został – w podobnym zapisie w tekstach mówionych nie było takiej konieczności, gdyż są one tekstami stworzonymi według pewnej specjalnej konwencji.

Teksty mówione to poza tym niejednokrotnie rozmowy kilku osób, które odzywały się jednocześnie, wchodziły sobie w słowo, przerywały wypowiedź, a przerwaną wznawiali – raz kontynuując myśl bez zmian, a innym razem odnosząc się do tego, co w międzyczasie powiedziano. Anotatorzy otrzymywali z tego zapisu powycinane losowo fragmenty. Niestety, mamy wrażenie, że dobra anotacja (i spójna z anotacją innych typów tekstów) takich rozmów jest zadaniem niezwykle trudnym. Otrzymujemy bowiem przybliżony, lecz nieco wykrzywiony obraz języka mówionego (a już na pewno na poziomach segmentacyjnych i składniowych). Być może należałoby zatem opracować osobny system anotacji takich tekstów.

Technikalia

Wydaje się, że Anotatornia bardzo dobrze spełniła swoje zadanie. Anotacja NKJP odbyła się bez większych problemów natury technicznej, co warte jest

[18] Można by się spodziewać raczej zapisu *wu wu wu*.

podkreślenia ze względu na implementację w jednym systemie wyszukanych mechanizmów zarządzania tekstami, anotatorami i konfliktami w znakowaniu. Z perspektywy czasu dostrzegamy też to, czego zabrakło w samym systemie.

Po pierwsze, zabrakło wyszukiwarki już zaanotowanych tekstów, zwłaszcza z uwagi na superanotatorów, którzy powinni dbać o spójność anotacji; opcja taka dla samych anotatorów mogłaby być zachętą do kopiowania rozwiązań, które wcale nie są poprawne. Na przykład *pół* mogło być na początku często anotowane błędnie, niezgodnie z instrukcją znakowania, jako num (Morfeusz w ogóle nie proponował dla tego segmentu interpretacji subst)[19]. Anotatorzy, zwłaszcza na początkowym etapie pracy, najczęściej popełniali błędy właśnie tam, gdzie system nie wyświetlał wśród propozycji tej poprawnej i trzeba było samodzielnie wprowadzić nowy tag. Gdyby anotator sprawdzał dotychczasową anotację, ze względu na wątpliwości w jednym ze swoich akapitów, mógłby się upewnić, że anotuje dobrze, choć faktycznie robiłby źle.

Po drugie, przydatna wydaje się opcja, która pozwalałaby blokować możliwość pobrania nadmiarowej liczby transz (w stosunku do rozpoczętych, ale niezakończonych). Czasem, gdy ktoś pobierał zbyt dużo transz na zapas (nie były one bowiem wszystkie dostępne od razu, lecz przekazywano je do anotacji sukcesywnie, tak że zdarzały się czasem parodniowe przestoje w pracy), jedyną opcją było ich odbieranie i zwrócenie do puli wolnych transz; takie działanie wzbudzało jednak kontrowersje.

Z powyższym wiąże się też techniczne rozwiązanie wymuszające komentowanie rozbieżności po trafieniu akapitu do osądu. Bez wątpienia wymóg skomentowania własnej decyzji zmuszał do refleksji, powołania się na instrukcję (wymuszał jej znajomość). Być może warto było wprowadzić limit, np. pięć nieskomentowanych transz, oznaczający, że nie można pobrać nowej transzy, jeśli ma się pięć w konflikcie, oczekujących na komentarz.

Wreszcie szczegółem technicznym, acz lingwistycznie już poważniejszą kwestią, było znakowanie segmentów typu %, $. W obecnej wersji Anotatornia nie pozwalała na zmianę anotacji segmentów interpunkcyjnych, a do takich Morfeusz zaliczał powyższe symbole. Tymczasem, zgodnie z instrukcją znakowania, powinny być one tagowane tak, jak odpowiadające im wyrazy (*procent* i *dolar*).

6.7. Podsumowanie i perspektywy

Do powstania zasad znakowania morfosyntaktycznego opisanych powyżej przyczyniło się – pośrednio i bezpośrednio – wiele osób. Jak już wspomniano, zasady

[19] Dopowiedzmy dla jasności: instrukcja znakowania przewidywała zarówno interpretację rzeczownikową (jako podstawową), jak i liczebnikową.

te są w dużej mierze oparte na pracach Zygmunta Saloniego i Janusza S. Bienia i wywodzą się bezpośrednio z zasad znakowania Korpusu IPI PAN opracowanych przez Marcina Wolińskiego i Adama Przepiórkowskiego, z udziałem Łukasza Dębowskiego i Elżbiety Hajnicz; radami służył także Zygmunt Saloni. Różnice między tamtymi wytycznymi i niniejszymi zasadami wynikają częściowo z istotnych różnic między informacjami dostępnymi w dwóch wersjach analizatora morfologicznego Morfeusz (zob. p. 6.1), a częściowo z kilkuletnich doświadczeń w wykorzystaniu Korpusu IPI PAN. Wiele rozwiązań szczegółowych zostało zaproponowanych przez anotatorów NKJP i zebranych przez dwoje „superanotatorów": Annę Czelakowską i Łukasza Szałkiewicza, który w trakcie anotowania był także odpowiedzialny za utrzymywanie spójnej instrukcji znakowania.

Oczywiście zasady te są często wynikiem kompromisu między rozważaniami teoretycznymi i wymaganiami praktycznymi; wiele rozwiązań ma charakter arbitralny, niektóre problemy zaś, na przykład dotyczące tradycyjnych kategorii czasu, trybu czy też jednostek wielowyrazowych, zostały przesunięte na dalsze poziomy znakowania (zob. rozdziały 8 i 9).

Zasady znakowania morfosyntaktycznego opisane w niniejszym rozdziale – odpowiednio rozwinięte i obejmujące wiele przypadków szczegółowych – obowiązywały anotatorów znakujących milionowy podkorpus NKJP (zob. rozdz. 5). Podkorpus ten posłużył do wytrenowania narzędzia PANTERA, wykorzystanego następnie do automatycznego oznakowania pełnego korpusu NKJP (zob. rozdz. 11). Wyniki tej anotacji są dostępne do na stronie `http://nkjp.pl/`, za pomocą przeszukiwarki Poliqarp (`http://nkjp.pl/poliqarp/`). Niestety, automatyczne znakowanie jest zawsze obciążone pewnym błędem; w wypadku obecnej wersji PANTER-y ok. 7% segmentów oznakowanych jest nie do końca zgodnie z decyzjami podjętymi przez anotatorów na podstawie zasad streszczonych powyżej (zob. p. 11.4).

Dalsze prace, prowadzone obecnie przede wszystkim w ramach europejskiego projektu CESAR / META-NET (`http://www.meta-net.eu/projects/cesar`), zmierzają do zmniejszenia tego procentu błędów. Istotna poprawa powinna zostać osiągnięta zarówno przez zwiększenie korpusu treningowego – do podkorpusu milionowego dołączony zostanie na nowo oznakowany[20] korpus *Słownika frekwencyjnego…* (Kurcz i in. 1990), jak i przez redukcję błędów znakowania w podkorpusie milionowym. Niemniej jednak żywimy nadzieję, że – mimo pewnych niedoskonałości – już teraz zasoby tekstowe opracowane zgodnie z opisanymi tutaj zasadami w ramach projektu NKJP okażą się przydatne w różnorodnych pracach leksykograficznych, lingwistycznych i informatycznych.

[20] Do anotacji ponownie wykorzystywane jest opisane w p. 6.6.1 narzędzie Anotatornia.

7

Anotacja sensami słów

Izabela Will

Każdy użytkownik polszczyzny zdaje sobie sprawę z wieloznaczności takich słów jak *pióro* („wyrostek na ciele ptaka" lub „narzędzie do pisania"), *bal* („zabawa" lub „kawałek drewna"), *zasada* („reguła" lub „substancja chemiczna"), *prowadzić* („wykonywać jakieś działania" lub „zmierzać w jakimś kierunku"). Niewielu jednak uświadamia sobie, że większość wyrazów, którymi posługujemy się na co dzień, ma więcej niż jedno znaczenie.

W terminologii semantycznej wieloznaczność określa się jako polisemię lub homonimię. Za homonimy uważa się dwa wyrazy, które mają taką samą formę, ale różne znaczenie, a często też inną etymologię i odmienny paradygmat. Podobieństwo formalne między nimi jest więc niejako sprawą przypadku. Wspomniany już wyraz *bal* w znaczeniu „zabawa" pochodzi z języka francuskiego, a wyraz *bal* w znaczeniu „kawałek drewna" to staropolskie słowo. Polisemię można zdefiniować jako wiele spokrewnionych ze sobą znaczeń jednego wyrazu. Im dłużej słowo istnieje w danym języku, tym większe jest prawdopodobieństwo, że wyodrębni się z niego nowe znaczenie, np. słowo *rączka* może oznaczać „małą rękę" lub „uchwyt". W wypadku polisemii istnieje relacja między poszczególnymi znaczeniami. Odnosząc się do przykładu *rączki*, można powiedzieć, że wyodrębnienie się nowego znaczenia nastąpiło na zasadzie podobieństwa funkcji – część ciała, podobnie jak część walizki, służy do chwytania. Trzeba jednak pamiętać, że zarówno *polisemia*, jak i *homonimia* są terminami językoznawczymi. Różnica między nimi jest w dużej mierze umowna i nie dotyczy płaszczyzny treści. Dobrze ilustruje to angielski przymiotnik *gay*, który może oznaczać zarówno „wesoły, pogodny", jak i „homoseksualny". Jak pisze Saeed (2009: 65) to ostatnie znaczenie wyodrębniło się z pierwszego całkiem niedawno, co przemawiałoby za uznaniem tego wyrazu za polisemiczny. Jednak dla większości użytkowników angielskiego, zwłaszcza młodszych, oba znaczenia są tak odmienne, że skłonni są uznać wyraz *gay* za przykład homonimii.

Wieloznaczność słów najłatwiej dostrzec w kontekście, np.[1]:

(7.1) *Album **zawiera** sześćdziesiąt pięć reprodukcji.*

(7.2) *W maju **zawarłam** umowę z firmą budowlaną.*

Użytkownik języka polskiego będzie wiedział, że czasownik użyty w obu zdaniach oddaje inną treść, ma inne znaczenie, inny sens. Może mieć jednak problem ze sformułowaniem, na czym polega różnica między tymi znaczeniami. Kiedy jednak ma do dyspozycji odpowiednie „etykiety" schematycznie określające różne znaczenia tego czasownika – „być częścią lub składnikiem czegoś" i „ustalić" – bez trudu przypisze pierwszą z nich do zdania (7.1), a drugą do zdania (7.2).

7.1. Konstruowanie słownika sensów

Tworzenie słownika sensów[2] wykorzystywanego w ręcznym znakowaniu sensami milionowego podkorpusu NKJP, a także będącego częścią narzędzia do automatycznego rozróżniania sensów słów, przypomina tworzenie „etykiet" dla słów wieloznacznych. Nie wystarczy w tym celu posłużyć się definicjami wyróżniającymi znaczenia poszczególnych słów wziętymi z dostępnych słowników języka polskiego, gdyż zazwyczaj są one zbyt szczegółowe dla celów automatycznej dezambiguacji. Każdy z sensów, czyli takich „etykiet", musi być na tyle szeroki, by objął swoim znaczeniem szereg bardziej szczegółowych znaczeń, z których każde odpowiada temu bardziej ogólnemu[3]. Jednocześnie poszczególne sensy danego słowa powinny się od siebie różnić na tyle, by przypisanie leksemu w konkretnej wypowiedzi do któregoś z nich było jak najbardziej jednoznaczne.

Konstruowanie słownika sensów polegało na wybraniu około stu wyrazów wieloznacznych spośród najczęściej występujących w pewnym korpusie i opracowaniu dla nich haseł. Za bazę referencyjną przy wyborze najczęściej stosowanych jednostek leksykalnych posłużyła lista frekwencyjna opracowana na podstawie względnie zrównoważonego podkorpusu Korpusu IPI PAN i zawierająca kilka tysięcy leksemów. Każdemu z wyrazów przypisana była klasa gramatyczna oraz liczba jego wystąpień, jak w poniższym przykładzie:

```
być verb 442224
ten adj 234864
```

[1] Wszystkie przykłady pochodzą z *Innego słownika języka polskiego PWN*, red. Mirosław Bańko, Warszawa 2000.

[2] Słowo *sens* jest używane w niniejszym rozdziale w znaczeniu etykiety, którą można przypisać grupie bardziej szczegółowych znaczeń danego słowa. Jeśli zatem pojawia się stwierdzenie, że dany leksem ma cztery sensy, oznacza to, że przypisano mu cztery takie „etykiety".

[3] Na podobnej zasadzie jak w hiponimii, gdzie np. słowo *pies* obejmuje szereg bardziej szczegółowych określeń, tj. *jamnik, dog, kundel* itp.

```
to subst 148403
mieć verb 106137
rok subst 81623
pan subst 80271
móc verb 73730
bardzo adv 52933
```

Z dostępnej listy należało wybrać tylko wyrazy wieloznaczne, a następnie ustalić – na podstawie słownika i danych korpusowych – jaką liczbę sensów należy wyodrębnić dla danego słowa.

Przypisanie sensów ograniczone było do klasy gramatycznej, która przypisana była najczęściej występującemu na liście frekwencyjnej leksemowi. Jeśli dana forma wyrazowa, np. *sam*, może być jednocześnie przymiotnikiem (np. *sami nie damy rady*), rzeczownikiem (*sam* jako sklep samoobsługowy) i częścią zaimka (*ten sam*), a na liście frekwencyjnej przy jej najczęstszym wystąpieniu pojawił się znacznik „adj" (przymiotnik), to wzięte zostały pod uwagę jedynie te sensy, które wiążą się z przymiotnikiem *sam*.

Opracowany na potrzeby tego projektu słownik sensów zawiera 106 leksemów[4], w tym 50 rzeczowników, 48 czasowników i 8 przymiotników.

7.2. Kryteria wyodrębniania sensów „zgrubnych"

Zasadniczą część projektu stanowiło wydzielenie tzw. sensów zgrubnych, a więc przypisanie każdemu leksemowi nie więcej niż sześciu sensów (tę najwyższą liczbę sensów ma tylko jeden czasownik *brać*). Do opracowania haseł wykorzystano *Inny słownik języka polskiego* (Bańko 2000; dalej jako: ISJP), a ściślej jego elektroniczną wersję[5] zawierającą 100 tysięcy haseł leksykalnych. Definicje zawarte w ISJP wyróżniały kilkanaście, a niekiedy nawet kilkadziesiąt sensów szczegółowych, które należało połączyć w bardziej ogólne sensy.

Przy wydzielaniu poszczególnych sensów brano pod uwagę:

1. prawdopodobieństwo ich wystąpienia,
2. wyraźne różnice między poszczególnymi sensami.

Czasownik *wygrać* może oznaczać zarówno „zwyciężyć" (np. *wygrać zawody*), jak i „grać melodię" (np. *wygrał melodię zwycięstwa*), ale prawdopodobieństwo jego wystąpienia w tym drugim znaczeniu jest dość małe – w próbce korpusu[6] występuje tylko 2 razy na 147 wystąpień form tego czasownika – i nie przewidziano dla

[4] Ta liczba nie ma szczególnego uzasadnienia. Pilotażowa wersja projektu zakładała przebadanie stu leksemów, a ostatecznie pojawiło się ich nieco więcej.

[5] *Multimedialny słownik szkolny PWN* (Bańko 2005).

[6] Próbka Korpusu IPI PAN zawierająca 30 milionów segmentów; `http://korpus.pl/`.

niego odrębnego sensu. Tym samym czasownik *wygrać*, jako mający tylko jeden sens, nie znalazł się w słowniku.

Najmniej problemów nastręczały leksemy będące homonimami, którym w słownikach przyporządkowano dwa odrębne hasła, np. *mina* 1: wyraz twarzy, *mina* 2: materiał wybuchowy. W takim wypadku każdy z sensów zgrubnych odpowiadał jednemu z pary homonimów. Największy problem stanowiły leksemy o bardzo ogólnym znaczeniu, takie jak czasownik *dać*, któremu w ISJP przypisano 47 różnych sensów. Ten sam czasownik w *Słowniku języka polskiego* pod redakcją Mieczysława Szymczaka (Szymczak 1994) ma przypisanych dziesięć sensów (osiem dla czasownika *dać / dawać* i dwa dla czasownika *dać się / dawać się*), co i tak dla konstruowanego słownika sensów jest zbyt szczegółowym rozróżnieniem.

Większą część słownika sensów stanowią czasowniki i rzeczowniki, choć na liście znalazło się też kilka przymiotników. Ponieważ klasa gramatyczna wyrazu ma decydujący wpływ na wyróżnianie sensów, poniżej omówiony zostanie sposób tworzenia haseł dla tych trzech klas.

7.3. Sensy rzeczowników

Stosunkowo czytelnym kryterium odróżniającym poszczególne sensy rzeczownika jest denotacja, czyli zakres nazwy, innymi słowy zbiór obiektów, do których można odnieść dany sens rzeczownika, np. słowo *rada* można odnieść zarówno do grupy ludzi, jak i do porady słownej. Przyporządkowując sens do określonego desygnatu, uniezależniamy jego wydzielenie od kontekstu.

Innym sposobem na wyróżnienie sensu rzeczownika jest odwołanie się do powszechnie stosowanej kolokacji lub związku frazeologicznego zawierającego ten rzeczownik, np. trzeci z sensów wydzielonych dla wspomnianej już *rady* – możliwość zrobienia czegoś – odwołuje się do frazy: *dawać radę (coś zrobić), nie dawać rady*. Jeden z sensów rzeczownika *punkt* – sposób – odnosi się do powszechnie stosowanego wyrażenia *punkt widzenia*. Przeszukanie próbki Korpusu IPI PAN wykazało, że na 1000 wyników ze słowem *punkt*, 110 stanowiło właśnie wyrażenie *punkt widzenia*.

W tym miejscu należy podkreślić, że sensu, rozumianego jako jedna z etykiet przypisanych do danego leksemu, nie należy utożsamiać ze znaczeniem. Z punktu widzenia systemu do rozróżniania sensów słów nie jest istotne, czy wieloznaczność jest inherentną cechą danego wyrazu, czy też wynikiem zaistnienia kontekstu (np. obecności dopełnienia), kolokacji lub związku frazeologicznego. System bada bowiem otoczenie danego wyrazu i między innymi na tej podstawie przypisuje mu odpowiednią etykietę (por. p. 12.3). Jeśli więc obok słowa *punkt* pojawi się rzeczownik *widzenia*, system automatycznie przypisze mu sens „sposób",

abstrahując od faktu, że nośnikiem znaczenia jest całe wyrażenie *punkt widzenia*, a nie tylko jeden z jego członów.

7.4. Sensy czasowników

Kryteria pomocne przy wyróżnianiu rzeczowników, takie jak denotacja czy związki frazeologiczne, nie sprawdzają się w odniesieniu do czasowników. Ich sens bowiem, bardziej niż sens rzeczowników, zależy od stojącego po nich dopełnienia. Różnicę między wyrażeniami *dać komuś książkę* i *dać komuś spokój* można wytłumaczyć, odnosząc się do cech semantycznych rzeczowników wchodzących w skład tych fraz. Słowo *książka* jest rzeczownikiem konkretnym, a *spokój* rzeczownikiem abstrakcyjnym, dlatego pierwsze z podanych wyrażeń jest konkretne (odnosi się do konkretnej czynności dawania, w której wykonawca czynności przekazuje odbiorcy czynności jakiś obiekt), a drugie metaforyczne (nie zachodzi konkretny akt przekazania czegoś). Można też mówić o rozszerzeniu semantycznym pierwszego znaczenia na drugie. Jednak już wyrażenia takie jak *dać drapaka*, *dać plamę* są bardziej idiomatyczne, a ich znaczenia nie oddaje żaden z wyróżnionych sensów czasownika *dać*.

Inny problem stanowią frazy, w których czasownik *dać* jest częścią predykatu złożonego (analitycznego), np. *dać rozkaz*, *dać odpowiedź*, *dać radę*. Jak podaje ISJP – *dać odpowiedź* znaczy to samo, co *odpowiedzieć*, a *dać gwarancję* znaczy to samo, co *zagwarantować*. W takim przypadku należałoby mówić o braku sensu czasownika, gdyż sam czasownik jest semantycznie pusty, a „ciężar znaczeniowy" w powyższych frazach jest przerzucony na rzeczownik. Słownik sensów nie przewidywał wyróżniania „sensów pustych", a ograniczenie przypisania danemu leksemowi nie więcej niż pięciu czy najwyżej sześciu sensów niwelowało możliwość potraktowania każdego z powyższych wyrażeń jako odrębnego sensu. Ostatecznie czasownikowi *dać* przypisano trzy sensy:

(7.3) przekazać, podać coś; postawić coś przed kimś; oddać
__Daj__ psu jeść i wyprowadź go na dwór.

(7.4) być możliwym do zrobienia; pozwolić komuś coś zrobić
Chciałbym odbić ten tekst na ksero i powiększyć, jeśli się __da__.

(7.5) przypuszczać
Kiedy robotnik wraca po nocnej pracy w fabryce, __dajmy__ na to, o ósmej godzinie, to nie ma siły zajmować się dziećmi.

Od razu pojawia się pytanie, czy słuszne jest połączenie w jeden zgrubny sens tak różnych zastosowań czasownika jak *dać książkę* i *dać buziaka*. Na to pytanie nie ma jednoznacznej odpowiedzi, a decyzja, by sens zdefiniowany w przykładzie (7.3) był sensem prymarnym, a jednocześnie obejmował „puste" sensy czasownika

dać, może wydać się nazbyt arbitralna. Kontrowersyjna pozostaje również kwestia przypisania do tego samego sensu takich wyrażeń idiomatycznych jak *dać drapaka*. Jednak w odniesieniu do wielu związków frazeologicznych zachodzi relacja między jednym z wydzielonych sensów czasownika i jego użyciem w danym wyrażeniu, np. sens czasownika *dać* w wyrażeniu *dać komuś popalić* odpowiada definicji podanej w przykładzie (7.4), a sens tego czasownika w wyrażeniu *dać komuś słowo* odpowiada definicji w przykładzie (7.3).

Kolejnym zagadnieniem związanym z przypisaniem sensu czasownikom jest istnienie w języku polskim par aspektowych czasowników[7], np. *wystąpić / występować*. Taka para zawiera dokonaną i niedokonaną formę czasownika, które znaczeniowo różnią się tylko cechą aspektu (Saloni 2001). Praktyką powszechnie stosowaną w słownikach języka polskiego jest umieszczanie pary aspektowej czasowników pod jednym hasłem. ISJP nie jest w tym względzie wyjątkiem, zatem przy haśle *dawać* znajdziemy jedynie odnośnik do hasła *dać*. Niekiedy nawet para będąca przykładem supletywizmu morfologicznego (np. *wziąć / brać*) stanowi jedno hasło. Przy tworzeniu słownika sensów czasowniki będące członami pary aspektowej traktowane były jednak jako dwa odrębne leksemy. Takie rozwiązanie wynika zarówno ze struktury danych, jak i z natury projektu. Na liście frekwencyjnej człony par aspektowych podane były jako osobne jednostki (nie zawsze bowiem dokonana i niedokonana forma czasownika miała tak samo wysoką frekwencję), a w korpusie NKJP człony pary aspektowej reprezentowane są przez różne lematy[8]. Człony par aspektowych stanowią odrębne hasła, nawet jeśli sensy wyróżnione dla formy dokonanej i niedokonanej danego czasownika całkowicie się pokrywają. Przykładem czasowników, których hasła wyglądają tak samo, jest para aspektowa: *zająć / zajmować*. Obu czasownikom przypisane są te same definicje i te same cztery sensy zgrubne:

(7.6) zapełnić przestrzeń/miejsce/czas; zarezerwować miejsce; znajdować się gdzieś
***Zajął** książkami całe biurko.*

(7.7) przyciągnąć uwagę, zainteresować, zaciekawić
*Książka tak go **zajęła**, że o mało co nie przegapił swojej stacji.*

(7.8) zacząć się palić
*Świece gasły na deszczu, zapalała je ciągle, osłaniała ręką, póki się dobrze nie **zajęły**.*

(7.9) zacząć coś robić lub kimś się opiekować
*Po urodzeniu dziecka **zajęła** się domem.*

[7] Para aspektowa rozumiana jest w tym rozdziale jako para czasowników nieróżniących się w części przedrostkowej, np. *nastąpić / następować*, ale nie *pisać / napisać*.

[8] Lemat to podstawowa (hasłowa) forma danego leksemu, np. *pies* jest lematem słowa *psa*.

W słowniku sensów można jednak znaleźć takie pary aspektowe, których człony mają nieco inne znaczenie, np. czasowniki stanowiące parę *wystąpić* / *występować* mają dwa wspólne sensy:

(7.10) pokazać się, wziąć udział; zaprezentować coś, zdarzyć się
***Występuje** pani stale w Teatrze Dramatycznym.*

(7.11) oddzielić się, wysunąć naprzód, wyjść poza pewien obszar
***Wystąp** na środek i opowiedz wszystkim, co się stało.*

Poza wymienionymi dwoma sensami czasownik *występować* (ale nie *wystąpić*) ma przypisany jeszcze jeden sens, zdefiniowany w (7.12).

(7.12) znajdować się gdzieś
*Witamina A **występuje** w wielu pokarmach.*

Jeszcze jedną kwestią związaną z czasownikami, którą należało rozstrzygnąć, przystępując do opracowywania haseł, była obecność partykuły *się* jako składnika wielu czasowników. Niejednokrotnie forma czasownika z partykułą *się* ma zupełnie inne znaczenie niż ta sama forma czasownika bez tej partykuły, np. *stać* i *stać się*. Dla form czasownikowych występujących ze słowem *się* nie tworzono w słowniku sensów odrębnych haseł, lecz umieszczano je pod wspólnym hasłem, np. *stać się* tworzy jeden z sensów hasła *stać* (zob. (7.16)).

(7.13) mieć pozycję pionową, znajdować się gdzieś; wykonywać czynności na stojąco
*Z przyczyn zdrowotnych matka nie może długo **stać**.*

(7.14) zatrzymać się lub przerwać jakąś czynność; nie pracować, być w bezruchu
***Stójcie**, nie tak szybko.*

(7.15) mieć dość pieniędzy, zachować się w jakiś sposób
*Autostrady są potrzebne, ale czy nas na nie **stać**?*

(7.16) stać się: zdarzyć się, mieć miejsce
*Wolę nie myśleć, co by się **stało**.*

7.5. Sensy przymiotników

Większość opracowanych dla przymiotników haseł odpowiada hasłom z ISJP. Tę wyjątkową zbieżność oraz brak problemów z wyróżnieniem sensów przymiotników można tłumaczyć dwojako. Przede wszystkim przymiotniki pełnią funkcję atrybutywną, czyli są wyrazami określającymi rzeczowniki. Ich pojawienie się – przed czy po – określonym typie rzeczownika determinuje ich znaczenie. Można zilustrować tę zasadę przykładem przymiotnika *wolny*, któremu przypisano następujące trzy sensy:

(7.17) dowolny, niezależny, nieograniczony zewnętrznymi przepisami, zarządzeniami, zwyczajami
(7.18) niezajęty
(7.19) powolny

Jeśli przymiotnik *wolny* pojawi się przed rzeczownikiem *łóżko*, to tylko w sensie (7.18), jeśli z rzeczownikiem *związek*, to tylko w sensie (7.17), a jeśli z rzeczownikiem *ruch*, to tylko w sensie (7.17) lub (7.19).

Kolejnym powodem tego, że wyróżnienie sensów przymiotników okazało się stosunkowo proste, jest ich ograniczona rola w związkach frazeologicznych. O ile niemal wszystkie obecne w słowniku czasowniki tworzą jakieś wyrażenia idiomatyczne, o tyle przymiotniki, nawet będąc częścią utartych kolokacji czy zwrotów, np. *wolny ptak* czy *wolna amerykanka*, nie tracą przypisanego im znaczenia. Przymiotnik *wolny* użyty w obu wspomnianych zwrotach nadal odpowiada sensowi zdefiniowanemu w przykładzie (7.17).

7.6. Wykorzystanie słownika w NKJP

Implementacja słownika do systemu automatycznego rozróżniania sensów słów była najlepszym testem na weryfikację jego przydatności. Jak pokazały rezultaty tego eksperymentu (por. rozdz. 12), sensy wydzielone dla większości leksemów wieloznacznych miały odniesienie do danych korpusowych. Jednak znalazły się też hasła, w których podział na sensy był zbyt szczegółowy jak na zawartość korpusu i potrzeby automatycznej dezambiguacji. W wypadku tych haseł niektóre z sensów były słabo reprezentowane lub nie wystąpiły w ogóle. Na przykład czasownik *następować* miał w słowniku dwa sensy:

(7.20) wejść na coś stopą, atakować
*Kazik **nastąpił** widać na szyszkę, bo krzyknął boleśnie.*
(7.21) zdarzyć się
*Śmierć **nastąpiła** w nocy na skutek udaru serca.*

Jak wykazała statystyka, wszystkie 137 wystąpień tego czasownika odnosiło się do drugiego z wyróżnionych sensów.

Powyższy przykład pokazuje, że tworzenie słownika sensów dokonuje się na innej zasadzie niż opracowywanie haseł do tradycyjnych słowników. Jednym z najistotniejszych kryteriów skłaniających do wyodrębnienia sensu jest frekwencja użycia wyrazu w danym znaczeniu.

Po przeprowadzeniu ręcznej anotacji okazało się również, że niektóre wyrazy miały za małą liczbę wyróżnionych sensów. Gdy anotator nie mógł dopasować znaczenia wyrazu w tekście do żadnego z sensów znajdujących się w słowniku, zwiększał liczbę sensów o jeden.

Na przykład wyraz *góra* ma w słowniku wyróżnione następujące cztery sensy:

(7.22) wzniesienie, stos rzeczy zgromadzonych razem
(7.23) część przedmiotu położona wyżej od innych części; przestrzeń znajdująca się nad jakimś punktem odniesienia
(7.24) w wyrażeniu „z góry": wcześniej, zawczasu
(7.25) przewaga, władza

W anotowanych tekstach pojawiło się dużo wyrażeń typu *w ciągu dwóch z górą dziesięcioleci*, w których *góra* ma znaczenie „ponad, przeszło". Zostały one oznaczone przez anotatorów jako niepasujące do żadnego z powyższych sensów. Zatem do wyróżnionych sensów *góry* należałoby dodać jeszcze jeden:

(7.26) ponad, przeszło

Na koniec należy wspomnieć o problemie związków frazeologicznych omówionym w p. 7.3. Ze względu na ograniczenia dotyczące liczby sensów, jakie mogły być przypisane danemu leksemowi, został on w trakcie tworzenia słownika potraktowany po macoszemu, a związki frazeologiczne albo przyłączano do jednego z sensów danego leksemu (np. *dać słowo* przypisano do jednego z sensów czasownika *dać* zdefiniowanego jako „przekazać, podać, postawić coś przed kimś, oddać"), albo wyodrębniano dla nich osobny sens (np. *punkt widzenia* stanowi odrębny sens rzeczownika *punkt* – „sposób"). Wydaje się jednak, że związki frazeologiczne mające najwyższą frekwencję powinny być wydzielone jako osobne sensy, nawet gdyby zwiększyło to liczbę sensów danego leksemu do kilkunastu. Z semantycznego punktu widzenia takie rozwiązanie może wydać się pewnym przekłamaniem, gdyż z definicji związku frazeologicznego wynika, że jego znaczenia nie da się wyprowadzić ze znaczeń i reguł łączenia składających się nań wyrazów (Polański 1993). Jednak system rozróżniania sensów słów działa na poziomie form wyrazowych, a nie na poziomie wyrażeń, zatem sens trzeba przypisać jednemu z elementów wchodzących w skład wyrażenia. Istotne jest przy tym, by forma wyrazowa, do której przypisany jest sens, była nośnikiem kategorii gramatycznej całego związku frazeologicznego, np. we frazie *dostać kosza* sens powinien być przypisany do czasownika *dostać* (a nie do rzeczownika *kosz*), gdyż cały ten związek pełni funkcję czasownika.

8

Anotacja składniowa

Katarzyna Głowińska

8.1. Wstęp

Anotacja składniowa Narodowego Korpusu Języka Polskiego polega na określeniu zakresu i opisie cech konstrukcji składniowych na dwóch poziomach: na poziomie słów składniowych (gdzie określa się klasę słowa i wartości kategorii morfologicznych) i na poziomie grup składniowych (gdzie określa się typ grupy oraz centrum składniowe i semantyczne). Anotacja NKJP nie ma zatem na celu opisania pełnej struktury zdania, lecz częściowej, za to możliwie najbardziej prawdopodobnej. Główne elementy koncepcji opisu składniowego NKJP zostały oparte na pracy Przepiórkowskiego (2008).

Podstawowe założenie, jakie zostało przyjęte, było takie, by oznaczać jedynie te słowa i grupy składniowe, które można było precyzyjnie wskazać i opisać, aby gramatyka powstała w trakcie projektu mogła być zastosowana do automatycznej anotacji pełnego korpusu NKJP. Dlatego też ustalając zasady opisu, wystrzegano się rozwiązywania problemów, z którymi wiązało się ryzyko niespójności opisu (wynikających z podejmowania arbitralnych decyzji przez anotatorów). Zrezygnowano na przykład z włączania w większe całości grup przyimkowych[1] oraz z ręcznego poprawiania form bazowych czasowników z *się*.

Rezygnacja z próby wskazania pełnych zależności w zdaniach NKJP wynikała także z braku słownika walencyjnego odpowiedniego do potrzeb przetwarzania automatycznego i do wymagań leksykalnych tak dużego korpusu. Jednak za analizą powierzchniową przemawiała możliwość wykorzystania parsera Spejd, który powstał z myślą o tego typu zadaniach.

[1] Wyjątek od tej reguły stanowiły konstrukcje elektywne (np. *jeden z wielu*, *nikt z nas*), ponieważ dało się je precyzyjnie opisać za pomocą reguł.

8.2. Słowa składniowe

Z powodów przedstawionych w rozdz. 6 podział na klasy gramatyczne na poziomie morfosyntaktycznym odznacza się dużym stopniem szczegółowości. Ponadto wśród kategorii gramatycznych brak jest czasu, trybu i zwrotności, gdyż są one właściwe jednostkom większym niż tak wydzielone segmenty.

Na poziomie składniowym zdecydowano się połączyć niektóre segmenty w większe jednostki tak, aby móc opisać analityczne formy czasu, trybu, stopniowania czy spójniki nieciągłe. Zrezygnowano przy tym z zasad ciągłości i rozłączności obowiązujących na niższym poziomie, co znaczy, że dopuszczalna jest sytuacja, by elementy jednej jednostki językowej nie znajdowały się w bezpośrednim sąsiedztwie (np. *niech tutaj przyjdzie*) i by jeden segment należał do dwóch jednostek (np. kropka kończąca zdanie należąca jednocześnie do skrótu z kropką). Poza tym z punktu widzenia składni nie było powodu, by odróżniać niektóre klasy, można więc było je połączyć (np. subst, ger i depr w jedną klasę Noun).

Nowe jednostki wyrazowe, które wprowadzono na poziomie składniowym, nazwano słowami składniowymi. Podział na słowa składniowe jest nieco bliższy tradycyjnemu podziałowi na części mowy niż podział zaprezentowany w rozdz. 6. Przywrócone zostały kategorie czasu i trybu. Dla odróżnienia od poprzedniego poziomu symbole klas gramatycznych dla słów składniowych zaczynają się wielką literą.

W tab. 8.1 przedstawiono tagset dla słów składniowych. Dla każdej klasy podane zostały właściwe im kategorie gramatyczne, nawiasem kwadratowym oznaczono kategorie opcjonalne. W tab. 8.2 znajdują się natomiast kategorie gramatyczne i ich wartości.

Przy przejściu z poziomu morfosyntaktycznego na poziom słów składniowych na liście klas gramatycznych nastąpiły takie oto zmiany:

1. Noun to nowa klasa gramatyczna, łącząca trzy klasy z poprzedniego poziomu: subst, depr i ger.
2. Verbfin to nowa klasa gramatyczna opisująca formy osobowe czasowników; w jej skład weszło pięć klas z poziomu morfosyntaktycznego: fin, bedzie, aglt, praet i impt.
3. Forma nieprzeszła (fin) weszła do klasy Verbfin z wartością czasu fut dla czasowników dokonanych, a pres dla niedokonanych.
4. Forma przyszła (bedzie) stała się częścią analitycznej formy czasu przyszłego dla czasowników niedokonanych lub formą czasu przyszłego czasownika *być*.
5. Aglutynant (aglt) został połączony z pseudomiesłowem lub czasownikiem typu *winien* w klasę Verbfin.

Tabela 8.1. Tagset dla słów składniowych

Symbol	Klasa	Kategorie gramatyczne
Noun	rzeczownik	liczba : przypadek : rodzaj : [aspekt] : [zwrotność] : [zanegowanie]
Num	liczebnik główny	liczba : przypadek : rodzaj : akomodacyjność
Numcol	liczebnik zbiorowy	liczba : przypadek : rodzaj : akomodacyjność
Adj	przymiotnik	liczba : przypadek : rodzaj : stopień
Adjc	przymiotnik predykatywny	
Adv	przysłówek	[stopień]
Ppron12	zaimek nietrzecioosobowy	liczba : przypadek : rodzaj : osoba : [akcentowość]
Ppron3	zaimek trzecioosobowy	liczba : przypadek : rodzaj : osoba : akcentowość : poprzyimkowość
Siebie	zaimek *siebie*	przypadek
Verbfin	forma osobowa	liczba : osoba : czas : tryb : aspekt : zwrotność : zanegowanie : [rodzaj]
Imps	bezosobnik	tryb : aspekt : zwrotność : zanegowanie
Inf	bezokolicznik	aspekt : zwrotność : zanegowanie
Pcon	imiesłów przysł. współcz.	aspekt : zwrotność : zanegowanie
Pant	imiesłów przysł. uprzedni	aspekt : zwrotność : zanegowanie
Pact	imiesłów przym. czynny	liczba : przypadek : rodzaj : aspekt : zwrotność : zanegowanie
Ppas	imiesłów przym. bierny	liczba : przypadek : rodzaj : aspekt : zwrotność : zanegowanie
Winien	leksemy *winien, powinien*	liczba : osoba : rodzaj : czas : tryb : aspekt : zwrotność : zanegowanie
Pred	predykatyw	czas : tryb : aspekt : zanegowanie
Prep	przyimek	przypadek : [wokaliczność]
Conj	spójnik współrzędny	[nieciągłość]
Comp	spójnik podrzędny	[nieciągłość]
Interj	wykrzyknik	
Qub	kublik	[wokaliczność]
Brev	skrót	wymaganie kropki : klasa/typ
Xxx	ciało obce	
Interp	interpunkcja	

Tabela 8.2. Kategorie gramatyczne dla słów składniowych

Kategoria	Wartości
Liczba	sg, pl
Przypadek	nom, gen, dat, acc, inst, loc, voc
Rodzaj	m1, m2, m3, f, n
Osoba	pri, sec, ter
Stopień	pos, com, sup
Czas	pres, past, fut
Tryb	ind, imp, cond
Zwrotność	refl, nrefl
Aspekt	imperf, perf
Zanegowanie	aff, neg
Akomodacyjność	congr, rec
Akcentowość	akc, nakc
Poprzyimkowość	npraep, praep
Wokaliczność	nwok, wok
Wymaganie kropki	pun, npun
Nieciągłość	discr, ndiscr
Klasa/typ skrótu	NOUN, ADJ, ADV, QUB, PREP, CONJ, VERB, PPAS, PACT, XXX, NG, PrepNG, AdjG, DisG

6. Pseudoimiesłów (praet) wszedł w skład klasy Verbfin o wartości czasu past lub o wartości fut, jeśli stanowił element analitycznej formy czasu przyszłego.
7. Rozkaźnik (impt) wszedł do klasy Verbfin z wartością trybu imp.
8. Przymiotnik przyprzymiotnikowy (adja) został uznany za część przymiotnika z łącznikiem (np. *polsko* w *polsko-angielski*).
9. Przymiotnik poprzyimkowy (adjp) wszedł w skład jednostek wielowyrazowych (zob. p. 8.2.1), np. *ludzku* w wyrażeniu *po ludzku*.
10. Burkinostka (burk) stała się elementem jednostek wielowyrazowych (zob. p. 8.2.1).

Wprowadzenie poziomu słów składniowych pozwoliło na opis następujących form i zjawisk:

1. formy analityczne czasowników:
   a) formy czasu przyszłego składające się z formy przyszłej (bedzie) i bezokolicznika (inf) lub pseudoimiesłowu (praet), np. *będzie śpiewał/śpiewać*;
   b) formy trybu rozkazującego: *niech* z formą nieprzeszłą (fin), np. *niech poczeka*;

   c) formy czasu przeszłego, przyszłego predykatywów i czasowników typu *powinien*, np. *można było, można będzie*;
   d) formy czasu teraźniejszego predykatywów, np. *warto jest*;
   e) formy trybu przypuszczającego dla czasowników właściwych i predykatywów, np. *napisałby, można by*;
2. opisowe formy stopnia wyższego i najwyższego przymiotników i przysłówków, np. *bardziej pracowity*;
3. czasowniki zwrotne[2], np. *opalać się*;
4. zanegowane formy czasowników osobowych, predykatywów, czasowników typu *winien*, bezosobników, bezokoliczników i imiesłowów;
5. wyrazy zawierające łącznik, które na poziomie morfosyntaktycznym stanowiły trzy jednostki, zostały uznane za jedno słowo składniowe, należące do tej samej klasy składniowej, co segment po łączniku, np. *polsko-japoński*, *Bachleda-Curuś*;
6. liczby porządkowe pisane jako liczby z kropką stały się jednym słowem składniowym należącym do klasy Adj (z wyjątkiem liczb stanowiących elementy wypunktowania);
7. cudzysłowy otaczające segment utworzyły słowo składniowe wraz z tym segmentem, np. *„po amerykańsku”*;
8. skróty;
9. jednostki wielowyrazowe.

Ostatnie dwa zagadnienia zostaną omówione szczegółowo w dalszej części niniejszego punktu.

Większość elementów, które pisane są łącznie z innymi formami wyrazowymi, zostało do nich dołączonych, tworząc słowa składniowe (np. *pisał* | *by, kupił* | *em*). Do wyjątków należą poprzyimkowe formy zaimka *-ń* (np. w wyrazie *na* | *ń*) oraz partykuły *-ż(e), -li*, które zachowały swoją odrębność, tzn. nie tworzą z innymi segmentami słów składniowych, mimo że od innych słów nie oddziela ich spacja.

Każdy segment z poziomu morfosyntaktycznego musiał otrzymać swój odpowiednik na poziomie słów składniowych. Przeważnie segmenty zostały przepisane na poziom wyższego rzędu, część z nich natomiast weszła w skład większych jednostek. W tab. 8.3 przedstawiono, ile słów składniowych składa się z jednego segmentu, a ile z dwóch i więcej segmentów.

Z tego zestawienia wynika, że 94% słów składniowych jest tworzonych przez jeden segment, ale nie oznacza to, że nie nastąpiła żadna zmiana poza zmianą symbolu klasy na pisany wielką literą. Były bowiem i takie przypadki, że nie

[2] Ustalono, że każdy element *się* powinien zostać dołączony do właściwego czasownika, nawet jeśli nie jest to czasownik zwrotny. Konsekwencją tej decyzji było dołączanie *się* do czasowników, które nie są zwrotne (np. w zdaniach *Szacuje się, że…*, *Powinno się przychodzić punktualnie.*).

Tabela 8.3. Struktura słów składniowych

Liczba segmentów	Liczba słów składniowych	Przykład
1	930 781	*wśród, liniowych, wchodziła*
2	52 090	*dopiero co, bać się*
3	10 269	*bądź co bądź, polsko-rosyjski*
4	481	*pożegnałbym się*
5	57	*bym się nie zdziwił*
6	0	–
7	4	*polsko-brytyjsko-holendersko-kanadyjski*

zmieniał się zakres, ale zmieniała się klasa gramatyczna i forma podstawowa (np. *jesienią* jako segment było rzeczownikiem, jako słowo składniowe w zdaniach typu *Przyjechała do nas jesienią*. stawało się przysłówkiem i zmieniało formę podstawową z *jesień* na *jesienią*), a czasem tylko jedna informacja morfologiczna (np. formy wołacza równego mianownikowi na poziomie morfoskładniowym opisywane były często jako formy mianownikowe; na poziomie składniowym zmieniano opis na wołaczowy).

8.2.1. Jednostki wielowyrazowe (JW)

Jednym ze zjawisk, które zostały opisane za pomocą słów składniowych, są jednostki wielowyrazowe (w skrócie JW). Termin ten, przyjęty na potrzeby projektu, obejmuje ustabilizowane połączenia wyrazowe, charakteryzujące się tym, że między składniki nie można wstawić żadnych wyrazów. Opisem objęto jednostki nieodmienne, pominięto więc połączenia wyrazowe zawierające rzeczowniki, czasowniki, przymiotniki i liczebniki.

Dzięki temu, że jednostki wielowyrazowe otrzymują status słowa składniowego, ich opis składniowy jest bliższy funkcji pełnionej w wypowiedzeniu niż ich budowie strukturalnej, np. *do czysta* strukturalnie to fraza przyimkowo-nominalna, która staje się przysłówkiem na poziomie słów składniowych i grupą przysłówkową na poziomie grup składniowych.

Wstępna lista zleksykalizowanych jednostek wielowyrazowych powstała na podstawie prac Milewskiej (2003a, 2003b) i Czerepowickiej (2005). Lista ta była rozszerzana w trakcie anotacji o jednostki opisane w *Innym słowniku języka polskiego* (Bańko 2000) jako połączenia zleksykalizowane.

Na liście jednostek wielowyrazowych znalazły się:

1. przysłówki, np. *po ludzku, do czysta, na czczo, ani trochę, co niemiara*;
2. kubliki, np. *bez mała, na pewno, bez wątpienia, bądź co bądź*;

3. przyimki złożone, np. *co do, w przeciwieństwie do, bez względu na*;
4. spójniki złożone, np. *a zatem, chyba że, podobnie jak, z tym że*;
5. spójniki nieciągłe, np. *nie tylko…, lecz…, zarówno…, jak i…, im…, tym…*

Sporym utrudnieniem w opisie regułowym tego typu połączeń wyrazowych była ich wieloznaczność. W jednym kontekście mogą one zostać uznane za jednostki wielowyrazowe, w innym nie. Oto dwa przykłady: *z prawa* (JW: *z prawa i z lewa*, grupa przyimkowo-nominalna: *korzystać z prawa do zasiłku*), *do reszty* (JW: *Do reszty utracił autorytet.*, grupa przyimkowo-nominalna: *Dołączył do reszty.*). W takich sytuacjach, tworząc regułę, starano się szczegółowo określić kontekst, w jakim mogło wystąpić zwykłe, a w jakim ustabilizowane połączenie. Jeśli było to niemożliwe, rezygnowano z opisu danej jednostki w gramatyce, a jej opis pozostawiono anotatorom.

Inny problem to pełnienie różnych funkcji w zdaniu przez daną jednostkę. Częste na przykład jest użycie wyrażeń przysłówkowych jako modyfikatorów rzeczowników, co pociąga za sobą konieczność zmiany klasy z przysłówka na przymiotnik, np. *na przełaj*: *Biegł na przełaj.* (Adv), *droga na przełaj* (Adj). Tego typu zmiany wprowadzane były przez anotatorów na podstawie analizy funkcji danej jednostki w zdaniu.

8.2.2. Skróty

Na poziomie morfosyntaktycznym skrótowce opisano jako rzeczowniki (subst), skróty zaś uznano za odrębną klasę (brev). Jako formę podstawową skrótu podawano jego rozwinięcie, nie określano jednak wartości kategorii gramatycznych.

Na poziomie składniowym pojawił się problem, jak traktować skróty, mając do dyspozycji jedynie dwie informacje: klasę i ich rozwinięcie. Rozważono trzy możliwości: 1. traktować skróty w sposób zbliżony do ich rozwinięć, a więc *płk* jak wyraz *pułkownik*, *br.* jak wyrażenie *bieżącego roku*; 2. włączyć skrót w obręb większej grupy składniowej, np. *k.k.* (*kodeks karny*) zwykle występuje po numerze paragrafu; 3. uznać skróty za odrębne jednostki składniowe i pozostawić je bez opisu składniowego. Zdecydowano się na rozwiązanie pierwsze, najbardziej intuicyjne. Pociągnęło to jednak za sobą konieczność wprowadzenia nowej kategorii gramatycznej o nazwie klasa/typ skrótu (brev_pos) oraz przygotowania reguł opisujących każdy skrót – dla skrótów, których rozwinięciem był pojedynczy wyraz, określano klasę na wzór klas słów składniowych (np. NOUN, ADJ[3]), a dla tych, których rozwinięciem była fraza, wskazywano typ grupy składniowej (np. NG,

[3] Ze względu na wymagania parsera klasa skrótu musiała różnić się od klas słów składniowych, pisana jest więc wersalikami.

PrepNG). Jeśli rozwinięciem skrótu była jednostka wielowyrazowa, to otrzymywał on opis jak odpowiednie słowo składniowe, a więc *np.* uznano za QUB, nie zaś PrepNG. W tab. 8.4 przedstawiono przykładowe skróty wraz z opisem (pun oznacza skrót wymagający kropki, npun skrót bez kropki).

Tabela 8.4. Opis skrótów na poziomie składniowym

Skrót	Rozwinięcie	Opis
w.	*wiek*	Brev:pun:NOUN
św.	*święty*	Brev:pun:ADJ
np.	*na przykład*	Brev:pun:QUB
m3	*metr sześcienny*	Brev:npun:NG
n.p.m.	*nad poziomem morza*	Brev:pun:PrepNG

Jak powiedziano wyżej, skrótom nie zostały przyporządkowane wartości kategorii typu przypadek, rodzaj. Nie można było ich więc uwzględnić w regułach opisujących grupy składniowe na równi z innymi jednostkami, ponieważ nie można było sprawdzić, czy np. skrót przymiotnika uzgadnia wartości z modyfikowanym rzeczownikiem. Powstał więc osobny zestaw reguł dla grup składniowych, w których mogą wystąpić skróty. Niosą one jednak ze sobą ryzyko złego dopasowania. Weźmy na przykład regułę dla grupy przyimkowo-nominalnej, na którą składa się przyimek i skrót, którego rozwinięciem jest rzeczownik lub grupa nominalna, typu *w br.*. Możemy podać elementy tej grupy: Prep i Brev, dla którego brev_pos ma wartość NOUN lub NG, nie możemy jednak nałożyć warunku, by przypadek wymagany przez przyimek zgadzał się z przypadkiem rzeczownika lub centrum grupy nominalnej.

Inna różnica między opisem skrótów i zwykłych wyrażeń polegała na tym, że w skrótach, których rozwinięciem było wyrażenie, centrum semantyczne i składniowe (zob. p. 8.3) było przypisane do skrótu jako całości (np. *n.p.m.*), natomiast w przypadku wyrażenia *nad poziomem morza* przyimek byłby opisany jako centrum składniowe, a pierwszy z rzeczowników jako centrum semantyczne.

Ponadto skróty wymagające kropki tworzą słowo składniowe wraz z następującą po nich kropką.

8.3. Grupy składniowe

Granica między słowami i grupami składniowymi nie jest wyraźna. Nie wiadomo na przykład, czym są ze składniowego punktu widzenia jednostki wielowyrazowe czy daty. Na potrzeby anotacji NKJP przyjęto, że jeśli w regule opisującej dany ciąg wyrazowy trzeba się odwołać do konkretnych form wyrazowych, to jest to

słowo składniowe. Jeśli zaś używa się głównie symboli klas gramatycznych, to jest to grupa składniowa. Tak więc jednostki wielowyrazowe stanowią w tym ujęciu słowa składniowe, ponieważ chcąc opisać np. kublik *na dodatek* trzeba podać w regule formy *na* i *dodatek*, a nie symbole klas gramatycznych: Prep i Noun.

Jak powiedziano na początku, ze względu na brak słownika walencyjnego, analiza składniowa Narodowego Korpusu Języka Polskiego polega na wskazaniu grup składniowych z pominięciem grup werbalnych. To oznacza również, że nie są łączone w jedną grupę dwa wyrazy, których zależność wynika z walencji, np. *miły sercu*, *winny braku nadzoru*.

Dla każdej grupy wskazuje się jej centrum składniowe i semantyczne. Centrum składniowe to ten element, który definiuje podstawowe własności dystrybucyjne danej grupy. Za centrum semantyczne zaś uznaje się ten składnik, który najlepiej definiuje znaczenie danej grupy. Na przykład w wyrażeniu przyimkowym *nad morzem* centrum składniowym jest przyimek, semantycznym zaś rzeczownik, natomiast w grupie liczebnikowej *pięciu chłopców* centrum składniowym jest liczebnik, a semantycznym rzeczownik. Argumenty przemawiające za wyróżnianiem obu centrów przedstawione są w pracy Przepiórkowskiego (2008).

W odróżnieniu od poziomu słów składniowych, gdzie jeden segment mógł być elementem dwóch słów składniowych (np. *się* w zdaniu *Bał się odezwać*.), ustalono, że jedno słowo składniowe nie może należeć do różnych grup składniowych.

W anotacji NKJP wyróżniono następujące grupy składniowe:

1. Grupa nominalna (NG); centrum składniowe i semantyczne stanowi rzeczownik (Noun), zaimek osobowy (Ppron12, Ppron3) lub zaimek *siebie*;
   a) grupa nominalna skoordynowana (*Jan albo Maria, rządu i parlamentu*); za centrum składniowe i semantyczne uznaje się pierwszy element grupy;
   b) z podrzędnikiem rzeczownikowym:
      i) oba rzeczowniki lub centra grup nominalnych w tym samym przypadku (*terroryści samobójcy*);
      ii) drugi rzeczownik lub centrum grupy nominalnej w dopełniaczu (*brat ojca, sala posiedzeń senatu, zabójca króla Henryka IV*);
      iii) rzadziej: drugi rzeczownik lub centrum grupy nominalnej w innym przypadku (*spacer ulicami Wrocławia*);
   c) z podrzędnikiem liczebnikowym:
      i) centrum składniowe grupy liczebnikowej w dopełniaczu (*kurtki trojga dzieci*);
      ii) rzadziej: centrum składniowe grupy liczebnikowej w innym przypadku (*spacer trzema ulicami Wrocławia*);
   d) z podrzędnikiem przymiotnikowym:

i) między przymiotnikiem i rzeczownikiem występuje uzgodnienie przypadka, liczby, rodzaju (*miła dziewczyna, bieżących wydarzeń politycznych*);
ii) rzeczownik z przymiotnikiem w dopełniaczu l.poj. rodzaju nijakiego (*coś dobrego, nic dziwnego*);

e) z podrzędnikiem o postaci kublika (*prawie geniusz, [przed] niespełna rokiem*).

2. Grupa liczebnikowa (NumG); centrum składniowe stanowi liczebnik (Num), centrum semantyczne stanowi rzeczownik lub odsłownik (Noun)[4];
   a) z podrzędnikiem rzeczownikowym:
      i) między liczebnikiem a rzeczownikiem lub centrum grupy nominalnej występuje uzgodnienie przypadka, liczby i rodzaju (*dwie dziewczyny*);
      ii) liczebnik narzuca przypadek rzeczownikowi lub centrum grupy nominalnej (*sześć samochodów, dwoje skrzypiec*);
   b) z podrzędnikiem przymiotnikowym: między przymiotnikiem a liczebnikiem lub między przymiotnikiem a rzeczownikiem występuje uzgodnienie przypadka, liczby i rodzaju (*ostatnie pięć minut*, *[Widzę] pozostałych/pozostałe siedem kart.*);
   c) z podrzędnikiem przyimkowym: rodzaj liczebnika zależy od rzeczownika (*dwóch z kolegów, dwie spośród koleżanek*);
   d) z podrzędnikiem o postaci kublika (*dokładnie trzy minuty, niespełna pięć lat, ledwie sto metrów*).
3. Grupa przymiotnikowa (AdjG); centrum składniowe i semantyczne stanowi przymiotnik (Adj);
   a) z podrzędnikiem czasownikowym (*[jest] gotowy zostać, [jest] przyzwyczajony pracować*);
   b) z podrzędnikiem o postaci:
      i) przysłówka (*wyjątkowo piękny, znacznie śmielsza*);
      ii) kublika (*dość głupi*).
4. Grupa przyimkowo-nominalna (PrepNG); centrum składniowe stanowi przyimek (Prep), centrum semantycznym zaś może być rzeczownik lub odsłownik (Noun), zaimek osobowy (Ppron12, Ppron3) bądź zaimek *siebie* (Siebie) (*nad głównym wejściem, ze mną, nad sobą*).
5. Grupa przyimkowo-przymiotnikowa (PrepAdjG); centrum składniowe stanowi przyimek (Prep), centrum semantycznym zaś jest przymiotnik

[4] Jeżeli grupa liczebnikowa składa się z samego liczebnika, stanowi on zarówno centrum składniowe, jak i semantyczne frazy.

(Adj) (*[wyglądać] na zmęczonego, [wyprzedzać] na trzeciego, [uznać kogoś] za winnego*).

6. Grupa przyimkowo-liczebnikowa (PrepNumG); centrum składniowe stanowi przyimek (Prep), centrum semantycznym zaś jest rzeczownik (Noun) (*z dwiema osobami*), a jeśli go nie ma, to liczebnik (Num) (*[pić] za trzech*).
7. Grupa przysłówkowa (AdvG); centrum składniowe i semantyczne stanowi przysłówek (Adv); z podrzędnikiem o postaci: przysłówka (*gdzieś daleko*) lub kublika (*niemal natychmiast, ledwie widocznie*).
8. Dyskurs (DisG) – są to takie elementy zdania, które nie są składniowo związane[5] (*a nuż, no cóż, itp. itd., np., m.in., moim zdaniem*).
9. Zdanie podrzędne z *że, żeby, iż, aby, by* (CG); centrum składniowe stanowi spójnik podrzędny (Comp), centrum semantycznym zaś jest centrum zdania podrzędnego (forma osobowa czasownika lub forma predykatywu).
10. Zdanie podrzędne pytajne (KG); centrum składniowe stanowi zaimek pytajny (*kto, kogo, co, czy, czym, kiedy, gdzie, jak, jaki*), centrum semantycznym zaś jest centrum zdania podrzędnego (forma osobowa czasownika lub forma predykatywu).

Jako jedna grupa składniowa oznaczany jest możliwie najdłuższy ciąg słów składniowych, który należy do jednego z wyżej wymienionych typów. Grupa taka może być jednoelementowa, np. grupa nominalna składająca się z rzeczownika, może być też rozbudowana, np. grupa przyimkowo-nominalna zawierająca w sobie grupę nominalną: *w warszawskim kościele św. Krzyża*, tzn. w jej skład może wchodzić kilka grup zagnieżdżonych (aby zbudować tę grupę, parser korzysta z czterech reguł). Owe grupy składowe nie są jednak widoczne podczas anotacji, opisuje się bowiem tylko największą grupę. Wyjątek stanowią grupy KG i CG, czyli zdania podrzędne, wewnątrz których zaznaczone są wszystkie grupy składniowe.

Może się jednak zdarzyć, że grupa o największym zakresie w rzeczywistości składa się z dwóch grup, z których każda pełni inną funkcję w zdaniu. Na przykład w zdaniu *Szukał w domu książki.* można wskazać jedną grupę przyimkowo-nominalną (*w domu książki*) lub – co bardziej prawdopodobne – dwie grupy: przyimkowo-nominalną (*w domu*) i nominalną (*książki*), które wypełniają dwie pozycje w zdaniu (w ujęciu tradycyjnym): okolicznika miejsca i dopełnienia.

Chcąc zachować spójność opisu, nie włączano w obręb większych całości grup przyimkowych (przyimkowo-nominalnych, przyimkowo-przymiotnikowych i przyimkowo-liczebnikowych), ponieważ w wielu przypadkach nie można stwierdzić jednoznacznie, czy stanowią one jedną frazę, czy dwie (np. w zdaniu *Kupiła mieszkanie w Warszawie* nie można powiedzieć, czy wyrażenie

[5] Grupa typu DisG odpowiada temu, co Świdziński (1996) nazywa *członem innym*. Człon inny to składnik zdania, „który nie jest składniowo związany i o który nie sposób z sensem spytać".

w Warszawie stanowi część frazy nominalnej *mieszkanie*, czy łączy się składniowo z czasownikiem). Dlatego też grupy nominalne, przymiotnikowe i przysłówkowe z podrzędnikiem przyimkowym traktowane są jako dwie grupy (kreską pionową zaznaczono granice między grupami): *ochota | na kawę*, *odporny | na zabrudzenia*, *spektakl | pt. „Dziady"*.

Wyjątek od tej reguły stanowią konstrukcje elektywne, które zawierają podrzędnik przyimkowy, a opisywane są jako pojedyncze grupy. Są to konstrukcje o schemacie: „X + *z* / *spośród* + rzeczownik / zaimek osobowy / przymiotnik / grupa nominalna". Centrum składniowe grupy stanowi element X i od jego charakterystyki morfologicznej zależy typ grupy. Tak więc: *dwóch, wielu z…* to grupy liczebnikowe, *jeden, niejeden, któryś, którykolwiek, niektórzy z…* to grupy przymiotnikowe, a *ktoś, ktokolwiek, cokolwiek, nikt, nic z…* to grupy nominalne. Centrum semantycznym jest główny element frazy występującej po przyimku.

Innymi (poza przyimkami) elementami zdania, które wpływają na podział grupy, są znaki interpunkcyjne. Słowa lub ciągi słów, które występują w nawiasach, po myślniku, przecinku czy w inny sposób oddzielone są przez znaki interpunkcyjne, traktowane są jako odrębne części tekstu, np. we fragmencie *członek rady zakładowej – Jan Mizerakiewicz* zostaną wyróżnione dwie grupy nominalne. Wyjątek stanowią cudzysłowy otaczające grupy składniowe, które mogą stać się częścią grupy, np. *w „The Washington Post"* (PrepNG).

Następnym elementem, który może wyznaczać granice frazy nominalnej, jest imiesłów przymiotnikowy. Jeśli występuje on przed rzeczownikiem, stanowiącym centrum grupy nominalnej, i jeśli między imiesłowem a rzeczownikiem jest uzgodnienie przypadka, liczby i rodzaju, np. *planowany wzrost przychodów*, to uznaje się go za składnik danej grupy nominalnej. Jeśli jednak występuje on po rzeczowniku, to traktuje się go jako element spoza grupy.

Osobnych grup natomiast nie stanowią skróty. Jak powiedziano wcześniej (zob. p. 8.2.2), skróty traktowane są jak inne wyrazy i wyrażenia, a więc np. *cm* uznaje się za grupę nominalną. Jeśli jednak skrót wchodzi w skład większej grupy, np. *100 cm* (NumG), to pełni takie same funkcje, jak wyraz stanowiący rozwinięcie skrótu. Jedynym problemem jest to, że nieraz centrum składniowe lub semantyczne trzeba przypisać do całego skrótu (np. w grupie nominalnej *br.* zarówno centrum składniowym, jak i semantycznym jest cały skrót), mimo że jego rozwinięcie stanowią dwa słowa i tylko jedno z nich należałoby uznać za centrum.

Pewnym odstępstwem od zasady, by anotować jedynie te słowa i grupy składniowe, które da się precyzyjnie wskazać oraz opisać za pomocą reguł dla parsera Spejd, była decyzja, by opisywać konstrukcje nieciągłe. Są to konstrukcje przerwane przez inny składnik, który do nich strukturalnie nie należy, np. *Skusili się na najwyższą od dwóch lat kumulację*.

W trakcie anotacji pojawiły się różne problemy, które nie zostały przewidziane na początkowym etapie projektu. Rozstrzygano je na liście dyskusyjnej. Jednym z takich zagadnień była anotacja pewnych ciągów wyrazowych, które nie są opisywane przez tradycyjne gramatyki, np. daty, adresy, godziny, numery telefonów, adresy WWW, wyniki meczów (np. *8:2*), czas jako wynik sportowy (np. *2:27.41 godz.*). Dla dat, adresów i godzin wprowadzono trzy specjalne grupy nominalne: NGdata dla dat, NGadres dla adresów i NGgodz dla zapisu godzin. Pozwoliło to na jednolity opis takich zjawisk niezależnie od tego, w jaki sposób sposób podano informację. Na przykład godzina 12.30 może być zapisana na kilka sposobów (w nawiasie podane są typy grup składniowych, jakie należałoby przypisać każdemu wyrażeniu): *12.30* (AdjG), *dwunasta trzydzieści* (AdjG), *godzina dwunasta trzydzieści* (NG), *wpół do pierwszej* (PrepAdjG). Numery telefonów, wyniki meczy, czas jako wynik sportowy opisuje się jako grupy liczebnikowe (NumG), natomiast adresy WWW, które na poziomie morfosyntaktycznym uznano za rzeczowniki (subst), stają się grupami nominalnymi.

Bardziej kłopotliwe okazały się nieoczekiwane połączenia składniowe, które nie są uznawane przez tradycyjne gramatyki. Pierwszym z nich jest połączenie przyimka z przysłówkiem. Większość takich wyrażeń można było uznać za jednostki wielowyrazowe (np. *po równo*, *za darmo*, *na czerwono*), a więc za jedno słowo składniowe. Podobnie opisano wyrażenie *od kiedy*, tj. jako jednostkę wielowyrazową, mimo że nią nie jest, aby nie tworzyć grupy przyimkowo-przysłówkowej. Innym nieoczekiwanym połączeniem jest rzeczownik, grupa przyimkowo-nominalna lub grupa liczebnikowa obok przysłówka (np. *chwilę później*, *w rok później*, *trzy dni wcześniej*). W tym wypadku uznano, że przysłówek jest nadrzędnikiem, a więc mamy do czynienia z grupą AdvG z nietypowym podrzędnikiem.

Jeszcze inny problem stanowiły zdania podrzędne (CG, KG), w których nie wystąpiła forma osobowa czasownika lub forma predykatywu (np. *Nie wiedział, dlaczego.*, *Myślę, że w jakiejś części na nas.*), ponieważ te formy uznaje się za centrum semantyczne zdania podrzędnego. W takich sytuacjach wybierano inny element zdania na centrum (zwykle ten, który był centrum składniowym), ponieważ każda grupa z zasady musi mieć wskazane oba centra.

Tabela 8.5 przedstawia liczbę składników tworzących grupy składniowe w ręcznie anotowanym korpusie milionowym (zob. rozdz. 5). Najwięcej grup składa się z jednego słowa składniowego (39%) bądź z dwóch słów (31%). Grupy zawierające od jednego do pięciu słów składniowych stanowią 96% grup. Najdłuższa grupa (składająca się z 68 słów składniowych) to zdanie podrzędne typu CG: *[Przypominam], że w orzeczeniu z dnia 20 października 1992 r. Trybunał Konstytucyjny stwierdził niezgodność postanowień przepisów art. 15 ust. 4 oraz art. 16 ust. 1*

i 2 w związku z ust. 3 ustawy z dnia 29 marca 1963 r. o cudzoziemcach w brzmieniu nadanym ustawą z dnia 19 września 1991 r. z art. 81 ust. 1 przepisów konstytucyjnych utrzymanych w mocy przez ustawę konstytucyjną z dnia 17 października 1992 r.

Tabela 8.5. Struktura grup składniowych

Liczba słów składniowych	Liczba grup składniowych	Bez zdań podrzędnych
1	119 300	119 295
2	95 156	95 101
3	49 354	49 248
4	19 924	19 548
5	8 890	8 321
6–10	9 540	6 495
11–20	2685	186
21–30	720	1
31–40	185	0
41–50	42	0
51–60	7	0
61–68	3	0

Interesujące jest również spojrzenie na długość grup składniowych bez zdań podrzędnych CG i KG (trzecia kolumna tabeli). Jak widać, w korpusie znalazło się pięć zdań podrzędnych jednoelementowych (zawierających jedynie wyraz *gdzie*, *dlaczego*, *kto*) oraz 55 zdań dwuelementowych (np. przecinek i wyraz *dlaczego* lub *dlaczego* i kropka kończąca zdanie). Najdłuższa grupa w tym zestawieniu to grupa nominalna, która zawiera 21 słów składniowych: *Struktura organizacyjna oraz ogólne przesłanki obowiązującego systemu finansowego i ewidencji księgowej zaopatrzenia produkcji i zbytu oraz ustalania i podziału wyników finansowych*.

8.4. Procedura

Wybrany fragment korpusu, liczący milion segmentów, który został wcześniej anotowany na poziomie morfosyntaktycznym (zob. rozdz. 6), podzielono na kilkanaście części. Każda partia tekstu była przetwarzana przez parser Spejd (Buczyński i Przepiórkowski 2009), dla którego przygotowano pierwszy zestaw reguł. Pliki ze zdaniami i proponowanym przez parser opisem były następnie sprawdzane i poprawiane przez anotatorów. W trakcie tej weryfikacji anotatorzy zgłaszali, w których miejscach parser dokonał złej analizy lub nie zaproponował właściwego opisu. Uwagi te wykorzystywano do przygotowania nowych reguł.

Każda następna partia korpusu była zatem analizowana w oparciu o ulepszoną wersję gramatyki. Ostateczna jej wersja została wykorzystana do automatycznej anotacji całego korpusu NKJP.

Wszelkie sugestie i pytania anotatorów wysyłane były na specjalnie stworzoną na potrzeby projektu listę dyskusyjną. Dzięki temu odpowiedź na pytanie wysłane przez jedną osobę docierała do wszystkich.

Weryfikacja anotacji proponowanej przez parser odbywała się lokalnie – w programie TrEd (zob. p. 8.4.3). Pliki do anotacji były pobierane z serwera[6], a po wykonaniu pracy tam odsyłane.

Każde zdanie było oglądane i modyfikowane przez dwóch anotatorów. Miejsca, w których wystąpiły niezgodności, zostały zweryfikowane przez superanotatora. Wyniki zgodności anotatorów ze sobą prezentuje tab. 8.6.

Tabela 8.6. Zgodność anotatorów

	Słowa składniowe	**Grupy składniowe**
Anotacja zgodna	980 687	276 164
Zbiór testowy	994 652	307 832
Zbiór wzorcowy	994 065	306 300
Dokładność	0,9859	0,8971
Pełność	0,9865	0,9016
F-measure	0,9861	0,8993

Dla każdych dwóch odpowiadających sobie plików ustala się, że jeden z nich jest „testowy", a drugi „wzorcowy". Obliczając dokładność, bierze się pod uwagę, ile testowych grup lub słów składniowych zostało znalezionych we wzorcu (muszą być zgodne klasy gramatyczne słów lub typy grup oraz muszą być te same zbiory elementów). Licząc pełność, sprawdza się, ile wzorcowych grup lub słów składniowych znajduje się wśród testowych. Ponieważ zbiór testowy i wzorcowy wybierane są w tym wypadku przypadkowo, bardziej interesująca jest miara F-measure, która uwzględnia zarówno dokładność, jak i pełność.

8.4.1. Gramatyka dla parsera Spejd

Gramatyka dla parsera Spejd tworzona była od podstaw. Wykorzystano w niej jedynie sześć reguł z pracy Przepiórkowskiego (2008). Obecnie składa się z 1187

[6] Pliki były pobierane z tzw. repozytorium SVN. SVN to skrót od *subversion*, czyli systemu kontroli wersji – narzędzia ułatwiającego m.in. archiwizację dokumentów, przywracanie poprzednich wersji plików oraz synchronizację pracy wielu osób nad jednym projektem.

reguł[7], z czego 350 reguł opisuje jednostki wielowyrazowe, 438 opisuje skróty[8], 124 opisuje pozostałe słowa składniowe, 275 zaś grupy składniowe.

Reguły ułożone są w określonej kolejności, ponieważ z reguł znajdujących się na początku korzystają reguły znajdujące się w dalszej części gramatyki. Na przykład, reguły opisujące grupy przyimkowo-nominalne odwołują się do wcześniejszych reguł opisujących różne typy grup nominalnych: z podrzędnikiem przymiotnikowym, rzeczownikowym, skoordynowane itp. Układ reguł w gramatyce jest następujący:

1. reguły opisujące jednostki wielowyrazowe;
2. reguły przypisujące skrótom wartości kategorii klasa/typ (zob. p. 8.2.2);
3. reguły dla słów składniowych, które mają skleić w jedno słowo dwa segmenty lub więcej;
4. reguły przepisujące pozostałe (pojedyncze) segmenty z poziomu morfosyntaktycznego na słowa składniowe;
5. reguły opisujące grupy składniowe zawierające określone słowa (np. *kościół św.* + rzeczownik pisany wielką literą w dopełniaczu, *uczeń kl.* + cyfra rzymska od I do VIII);
6. reguły opisujące proste grupy składniowe;
7. reguły opisujące grupy składniowe składające się z poprzednio opisanych prostych grup składniowych;
8. reguły klasyfikujące pozostałe (pojedyncze) słowa składniowe do odpowiednich grup składniowych;
9. reguły dla zdań podrzędnych.

Reguły były tworzone w taki sposób, by uniknąć nadmiernego dopasowania. Tak więc na formy, które były potencjalnie dwuznaczne, nakładano określone warunki, np. chcąc odnaleźć zleksykalizowany przysłówek *na stałe*, trzeba było sprawdzić, czy *stałe* to przymiotnik poprzyimkowy (adjp), nie zaś zwykły przymiotnik (adj), aby uniknąć oznaczenia jako przysłówka fragmentu zdania *Ta sala przeznaczona jest <u>na stałe</u> ekspozycje.* Dzięki takiemu podejściu możliwe było wykrycie błędów opisu na poziomie morfosyntaktycznym, gdy np. jako formę podstawową skrótu *br.* raz podawano *bieżącego roku*, a innym razem *bieżący rok*.

W gramatyce dla parsera podział na grupy, opisany w części p. 8.3 tego rozdziału, jest bardziej szczegółowy. U podstaw tej decyzji leżało przekonanie, że nie wszystkie typy grup nominalnych (w sensie strukturalnym) mogą wchodzić w skład większych grup na tych samych zasadach. I tak np. grupa nominalna (NG) została rozbita na kilka podtypów:

[7] Gramatyka wciąż jest rozwijana. Podane liczby to dane z marca 2011.

[8] Należy pamiętać, że jedna reguła opisująca skróty lub jednostki wielowyrazowe może dotyczyć od jednej lub kilkunastu jednostek językowych.

NGa – grupa, w której modyfikatorem rzeczownika jest przymiotnik;
NGg – grupa, w której podrzędnikiem rzeczownika stanowiącego centrum jest rzeczownik w dopełniaczu;
NGs – grupa, w której podrzędnikiem rzeczownika stanowiącego centrum jest rzeczownik w tym samym przypadku (apozycja);
NGk – grupa skoordynowana;
NGe – konstrukcja elektywna;
NGn – grupa zawierająca liczebnik w pozycji podrzędnika, np. *zdaniem wielu ekologów*;
NGb – grupa nominalna, której centrum jest skrót;
NGx – grupa składająca się z zaimka i przymiotnika w dopełniaczu, np. *nic nadzwyczajnego*;
NGc – grupa nominalna w cudzysłowie.

Spośród tych grup najbardziej wyróżniają się grupy NGb, zawierające skróty, które rzadko wchodziły w skład innych reguł, ponieważ, z powodu braku określonych informacji morfologicznych (zob. p. 8.2.2), niemożliwe było sprawdzenie, czy między składnikami grupy następuje uzgodnienie np. przypadka. Innym przykładem na to, że takie rozbicie grup jest potrzebne, może być reguła dla wyrażeń o schemacie X + *i inni, i inne*, gdzie w miejscu X nie jest dopuszczona grupa NGx.

8.4.2. Anotatorzy i superanotatorzy

Głównym zadaniem anotatorów było sprawdzanie, czy wygenerowana przez parser Spejd anotacja jest właściwa, a w szczególności: 1. czy wszystkie elementy zdania mają swój odpowiednik na poziomie słów składniowych; 2. czy opis morfologiczny słów składniowych jest poprawny (wyrywkowo); 3. czy wyrażenia, które można uznać za jednostki wielowyrazowe (zob. p. 8.2.1) zostały opisane jako słowa składniowe; 4. czy zakresy i opisy słów składniowych są właściwe; 5. czy zakresy i typy grup składniowych są właściwe.

Jeśli anotator stwierdził, że opis morfoskładniowy segmentu wydaje się niewłaściwy, wysyłał pytanie na listę dyskusyjną i, po uzyskaniu od superanotatora poziomu morfosyntaktycznego zgody, mógł dokonać zmiany. Poprawki te były zapisywane przez program TrEd (zob. p. 8.4.3), aby umożliwić naniesienie ich na pliki zawierające anotację morfosyntaktyczną w sposób automatyczny.

Ponadto anotatorzy proszeni byli o zgłaszanie wyrażeń zleksykalizowanych, które można było opisać jako słowa składniowe, a także o wskazywanie fragmentów zdań, dla których parser zaproponował niewłaściwy opis. Dzięki tym zgłoszeniom gramatyka dla parsera była stale ulepszana.

Zadaniem superanotatorów było porównywanie par plików, zawierających ten sam materiał wyjściowy, a opracowanych przez dwóch różnych anotatorów.

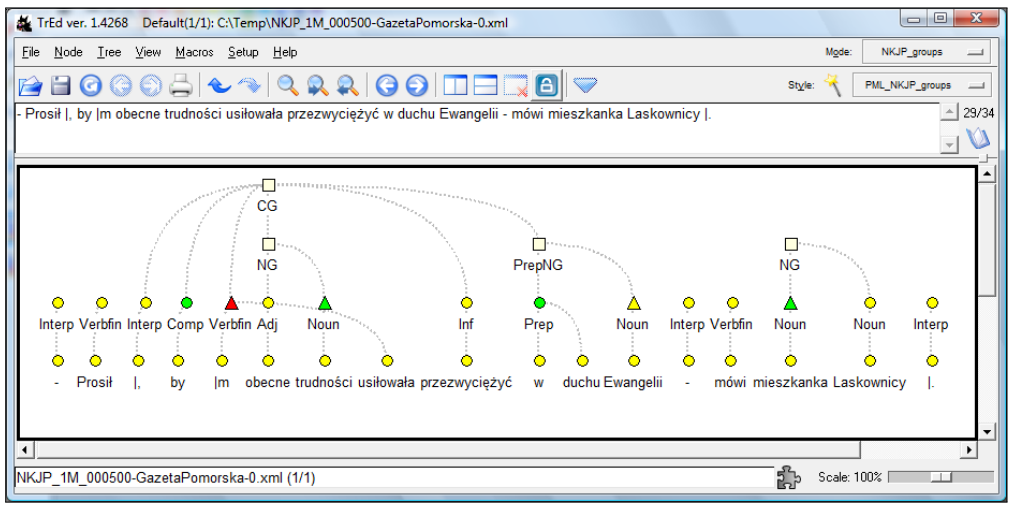
W razie stwierdzenia rozbieżności w anotacji superanotator dokonywał wyboru właściwego opisu.

8.4.3. Program TrEd

Anotacja poziomu składniowego, podobnie jak bytami nazwanymi, odbywała się w programie TrEd (`http://ufal.mff.cuni.cz/~pajas/tred`)[9], który został dostosowany do anotacji NKJP przez Jakuba Waszczuka.

Rysunek 8.1 przedstawia widok przykładowego zdania w programie TrEd. W głównej części okna znajduje się zdanie podzielone na segmenty (najniższy

Rysunek 8.1. Tred



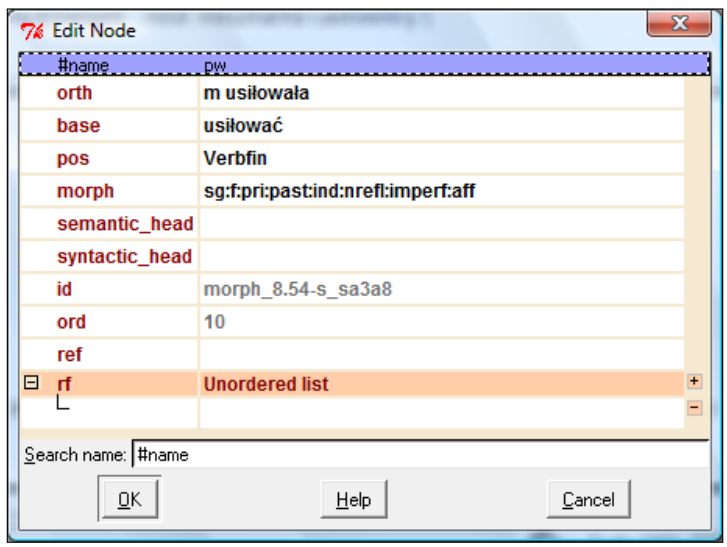
poziom). Każdy segment ma swój wierzchołek (żółta kropka). Ponad segmentami znajduje się poziom słów składniowych (również żółte kropki). Przy wierzchołkach podane są symbole klas gramatycznych (pisane wielką literą), do jakich należą te słowa. Jak widać, jedno słowo składniowe może łączyć dwa segmenty, które nie leżą w bezpośrednim sąsiedztwie (np. *m* i *usiłowała*). Na samej górze tej struktury znajdują się grupy składniowe (białe kwadraty). Przy wierzchołkach podany jest typ grupy. Słowo składniowe, które jest centrum składniowym (zob. p. 8.3) grupy, ma zielony wierzchołek, to zaś, które jest centrum semantycznym, ma wierzchołek w kształcie trójkąta. Powyżej głównej części okna podana jest treść całego zdania.

[9] W projekcie korzystano ze starszej wersji programu niż ta, która jest obecnie dostępna na stronie programu. Niestety, przygotowane na potrzeby projektu makra nie działają właściwie z najnowszą wersją programu.

Dzięki specjalnie przygotowanym makrom anotatorzy, podczas weryfikacji wyników automatycznego przyporządkowania znaczników przez parser, mogli wykonywać następujące operacje:

1. obejrzeć i w razie potrzeby zmodyfikować informacje morfosyntaktyczne dotyczące wybranej jednostki – zarówno z poziomu składniowego, jak i z poziomu segmentów podstawowych; informacje te mogły być wyświetlane w panelu bocznym lub w dodatkowym okienku (rys. 8.2), które pojawiało się po dwukrotnym kliknięciu wierzchołka;

Rysunek 8.2. Informacje o słowie składniowym



2. połączyć segmenty w słowo składniowe, wskazać jego klasę i wartości kategorii morfologicznych;
3. połączyć słowa składniowe w grupę (zaznaczając ciąg sąsiadujących ze sobą wierzchołków lub wybrane wierzchołki) i określić jej typ;
4. usunąć słowo składniowe lub grupę (jeśli parser niewłaściwie połączył elementy);
5. wyłączyć element z obrębu jednej grupy i włączyć go do innej grupy;
6. opisać jeden segment jako należący do dwóch słów składniowych (za pomocą dodatkowej krawędzi), np. *się* w zdaniu *Bał się odezwać.* jest elementem obu czasowników;
7. wskazać centrum składniowe i semantyczne grupy.

Z opcji dostępnych w oryginalnej wersji programu warto wspomnieć o możliwości wyszukania wierzchołka o określonej charakterystyce, np. danego słowa, grupy, segmentu o danej formie podstawowej czy opisie morfologicznym, a także

o możliwości przechodzenia do określonego zdania z pliku – ich lista dostępna jest w prawym górnym rogu po kliknięciu ikony książki.

Podczas zamykania pliku program sprawdza, czy każda grupa składniowa ma wskazane centrum składniowe i semantyczne, a także czy każdy segment ma swój odpowiednik na poziomie słów składniowych.

Superanotorzy również mieli do dyspozycji specjalnie przygotowane makra. Umożliwiały one przede wszystkim automatyczne wyszukiwanie różnic w opisie i przechodzenie od jednej różnicy do drugiej bez konieczności oglądania wszystkich zdań i ich elementów. Jeśli jeden z opisów był poprawny, można go było skopiować do wybranego pliku za pomocą skrótu klawiaturowego.

8.5. Podsumowanie

Anotacją składniową objęto ponad milion segmentów wchodzących w skład tej części korpusu, która była anotowana ręcznie, a właściwie półautomatycznie. W tab. 8.7 przedstawiono informacje dotyczące liczby różnych jednostek w korpusie. 67% z tej liczby słów składniowych weszło w skład grup składniowych, pozostałe 37% pozostało bez opisu na poziomie grup.

Tabela 8.7. Liczba elementów korpusu anotowanego składniowo

Element	Liczba
Segment	1 068 035
Zdanie	72 944
Słowo składniowe	993 684
Grupa składniowa	305 806

Tabela 8.8. Typy grup składniowych

Typ grupy	Liczba grup
NG	139 887
PrepNG	89 812
AdvG	32 014
AdjG	18 169
NumG	7 977
CG	6 099
DisG	3 873
PrepAdjG	3 235
PrepNumG	3 229
KG	1 511

Tabela 8.8 przedstawia liczbę wystąpień poszczególnych typów grup składniowych w korpusie milionowym, tab. 8.9 zaś liczbę wystąpień różnych słów składniowych.

Tabela 8.9. Klasy słów składniowych

Klasa gramatyczna	**Liczba słów składniowych**
Noun	297 904
Interp	177 081
Adj	113 006
Verbfin	99 980
Prep	97 957
Conj	37 110
Adv	36 988
Qub	27 785
Inf	16 282
Comp	15 539
Num	14 050
Ppas	12 013
Ppron3	12 000
Brev	9 971
Ppron12	6 335
Pred	5 288
Pact	5 091
Pcon	2 374
Imps	1 955
Siebie	1 812
Interj	1 368
Xxx	746
Winien	699
Pant	150
Numcol	125
Adjc	76

9

Anotacja jednostek nazewniczych

Agata Savary, Marta Chojnacka-Kuraś, Anna Wesołek,
Danuta Skowrońska, Paweł Śliwiński

9.1. Nazwy własne w systemie leksykalnym polszczyzny

Nazwy własne zajmują w systemie leksykalnym każdego języka miejsce szczególne. Zwykle przeciwstawia się je nazwom pospolitym, a podstawą takiego zestawienia jest funkcja obu typów nazw. Nazwy własne odnoszą się zwykle do jednostek, podczas gdy nazwy pospolite przysługują całym klasom jednostek. Funkcja indywidualizująca jednostek nazewniczych polega na „wyróżnieniu konkretnej osoby lub jednego obiektu spośród wszystkich takich samych lub podobnych obiektów danej klasy" (Rzetelska-Feleszko 2001b: 407). Mówiąc inaczej, nazwy własne – w przeciwieństwie do pospolitych – nie znaczą, nie mają konotacji w sensie logicznym; one po prostu nazywają (por. Urbańczyk 1992, Polański 1993).

Odwołanie się do pojęć treści i zakresu wyrazu pozwala stwierdzić, że wyrazy pospolite mają bardzo szeroki zakres, ale dość ubogą treść, za to nazwy własne mają zwykle bardzo wąski zakres, ale treść mogą mieć wyjątkowo bogatą. I tak wyraz *stół* odnosi się do wszystkich stołów, tj. przedmiotów, które możemy zaliczyć do tej klasy na podstawie przysługujących im cech. Natomiast treść tego wyrazu charakteryzuje się dużym stopniem ogólnikowości, gdyż może odnosić się do stołów różnego koloru, kształtu i różnej wielkości. Z kolei zakres nazwy własnej *Jan „Ptaszyn" Wróblewski* jest bardzo wąski, gdyż ograniczony do jednego, konkretnego człowieka. Zaś treść tej nazwy będą stanowiły wszystkie cechy kojarzone z daną osobą, noszącą to imię, pseudonim i nazwisko (por. Rzetelska-Feleszko 2001b, Urbańczyk 1992). Przypomnijmy jednak, że przy tego typu próbach zdefiniowania pojęcia nazwy własnej natrafiamy na znane problemy dotyczące hipotetycznej jednostkowości nazywanego obiektu. Dla przykładu pojedyncze imiona czy nazwiska, a także połączenia imion i nazwisk (np. *Jan*

Kowalski) mogą nazywać nie jednego, ale wielu osobników. Podobnie dzieje się w wypadku niektórych nazw geograficznych (np. *London*).

Nazwy własne od dawna są przedmiotem zainteresowania językoznawców, literatura przedmiotu z zakresu onomastyki jest ogromna – wykaz ważniejszych publikacji znajduje się m.in. w encyklopedii pod redakcją Stanisława Urbańczyka (1992). Mamy świadomość złożoności tej problematyki oraz wielości poruszanych w jej obrębie zagadnień, takich jak typologia nazw własnych, ich gramatyka, pochodzenie, specyfika regionalna (por. Rzetelska-Feleszko 2001a, 2005, Kosyl 2001). Jednak na potrzeby tego tekstu ograniczyliśmy się do wskazania podstawowych właściwości nazw własnych jako elementów systemu leksykalnego polszczyzny.

Od kilkunastu lat jednostki nazewnicze są również przedmiotem uwagi badaczy i praktyków z dziedziny przetwarzania języków naturalnych. Są one m.in. dobrymi hasłami do indeksacji i klasyfikacji tekstów, a także do ich automatycznego streszczania. Nazwy własne rządzą się specyficznymi prawami w procesie tłumaczenia, stanowią podstawowy element warstwy informacyjnej tekstów itd. Nie dziwi zatem, że jednostkom nazewniczym poświęcono osobny poziom opisu w projekcie anotacji Narodowego Korpusu Języka Polskiego (dalej jako NKJP).

9.2. Jednostki nazewnicze w polskiej leksykografii oraz światowej i polskiej lingwistyce korpusowej

Nazwom własnym (zwłaszcza osobowym i geograficznym) poświęcili wiele uwagi również polscy leksykografowie. Szczególne zasługi w tej dziedzinie mają Rospond (1992–1994) i Rymut (1980, 1992–1994, 1995, 2002, 2008). Z najnowszych opracowań warto wspomnieć o słowniku Rymuta (2002), zawierającym współcześnie używane w Polsce nazwiska wraz z liczbą ich wystąpień w poszczególnych powiatach. Inny słownik pod redakcją tego badacza (Rymut 2008) przedstawia opis hydronimów, tj. nazw rzek, kanałów, jezior i innych zbiorników wodnych w Polsce, z wykorzystaniem danych geograficznych, etymologicznych i historycznych. Z kolei leksykon pod redakcją Kubiak-Sokół i Łazińskiego (2007) zawiera około 7400 wybranych nazw polskich miejscowości wraz z odpowiadającymi im przymiotnikami relacyjnymi oraz nazwami mieszkańców w rodzaju męskim i żeńskim.

Anotacja jednostek nazewniczych w korpusach językowych również ma swoją tradycję w światowej lingwistyce komputerowej. Do jej popularyzacji przyczyniły się w szczególności konferencje MUC (Chinchor 1997), CoNLL-2002 i CoNLL-2003 (Sang i de Meulder 2003), umożliwiające konfrontację istniejących systemów do automatycznego rozpoznawania jednostek nazewniczych. Anotowane korpusy tworzone w ramach tych konferencji ukierunkowane były

przede wszystkim na ewaluację wydajności, z jaką konkurujące systemy potrafiły przypisywać uproszczony zbiór etykiet do pojedynczych segmentów korpusu. Na przykład korpus CoNLL-2003 ma postać przedstawioną na rys. 9.1, gdzie pierwsza kolumna zawiera kolejne segmenty korpusu, druga – etykiety morfosyntaktyczne, trzecia – znaczniki grup składniowych, a czwarta – znaczniki jednostek nazewniczych. Te ostatnie przybierają wartość `O` dla segmentu nienależącego do żadnej jednostki nazewniczej lub też `I-PER`, `I-ORG`, `I-LOC` i `I-MISC` dla elementów będących częściami składowymi jednostek nazewniczych (odpowiednio: nazw osób, organizacji, miejsc i inych nazw, takich jak wyrażenia numeryczne czy marki produktów).

Rysunek 9.1. Struktura korpusu CoNLL-2003

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC
.	.	O	O

Korpusy o podobnej strukturze zostały stworzone dla wielu innych języków. Głównym powodem popularności takich działań wydaje się fakt, iż korpusy takie mają w pierwszej kolejności służyć konkretnemu zastosowaniu, jakim jest automatyczna identyfikacja i klasyfikacja jednostek nazewniczych przy użyciu metod uczenia maszynowego (zob. też rozdz. 13).

Dużo rzadziej natrafić można na bardziej kompleksowe podejście, w którym jednostki nazewnicze są tylko jednym z aspektów lingwistycznej anotacji tekstów. Przykładem jest Praski Korpus Zależnościowy (Böhmová i in. 2003), tzn. korpus języka czeskiego z trzema warstwami: morfologiczną, składniową i „tektogramatyczną" (wyrażającą zależności semantyczne). Jednostki nazewnicze są w nim identyfikowane w warstwie tektogramatycznej wraz z innymi jednostkami wielowyrazowymi (Bejček i Straňák 2010), a następnie przypisywane do jednego z 9 typów i oznaczane odniesieniami do zewnętrznych zasobów leksykalno-semantycznych. Podobnie wielopoziomowy korpus języka holenderskiego (Desmet i Hoste 2010) zawiera milion słów z ręcznie poprawianą anotacją. Jedną z sześciu warstw semantycznych jest tworzona obecnie warstwa jednostek nazewniczych, które klasyfikowane są zgodnie z taksonomią liczącą 6 typów głównych i 17 podtypów. Ze szczególną uwagą traktowany jest tu problem metonimii (dla użyć metonimicznych podawane są informacje o typie zarówno „pierwotnym", jak i docelowym). Z kolei TüBa-DZ Treebank (Hinrichs i in. 2005)

jest wielopoziomowym korpusem języka niemieckiego, w którym jednostki nazewnicze są oznaczane na tym samym poziomie, co grupy składniowe, jednak (jak się wydaje) bez szczegółowej kategoryzacji.

Identyfikacja jednostek nazewniczych stała się częścią szeroko pojętej ekstrakcji informacji m.in. na forum ACE (Automatic Content Extraction)[1]. Korpusy używane tam w kampaniach ewaluacyjnych zawierają anotacje już nie tylko nazw własnych i spokrewnionych z nimi jednostek, ale również tak zwanych wzmianek (ang. *mentions*), czyli odniesień do obiektów świata rzeczywistego niezależnie od ich lingwistycznej postaci. Obok nazw własnych mogą tu zatem występować nazwy pospolite i różnego typu wyrażenia opisowe (np. *żona prezydenta*). Do innych ważnych cech tych korpusów należy uwzględnienie relacji zachodzących między anotowanymi jednostkami, jak również szczegółowa anotacja wyrażeń czasowych (absolutnych i względnych) oraz wydarzeń.

Wsród polskich korpusów anotowanych pod względem jednostek nazewniczych, poza NKJP, na uwagę zasługują m.in.: 1) korpus dialogów z infolinii warszawskiego Zarządu Transportu Miejskiego (Mykowiecka i in. 2008), zawierający ok. 81 000 słów i ok. 6200 oznaczonych jednostek nazewniczych, 2) korpusy ekonomiczny i giełdowy (Marcińczuk i Piasecki 2011) mające strukturę zbliżoną do pokazanej na rys. 9.1, zawierające ok. 330 000 słów i ok. 9000 oznaczeń nazw osób, miejsc i instytucji.

9.3. Modelowanie i metodologia anotacji nazw w projekcie NKJP

Z uwagi na rozmiar zadania, jakim jest ręczna anotacja korpusu zawierającego milion słów, za szczególnie ważne należało uznać jak najściślejsze określenie zakresu anotacji, a także reguł i strategii opisu każdej zidentyfikowanej w tekście jednostki.

9.3.1. Zakres anotacji i hierarchia nazw

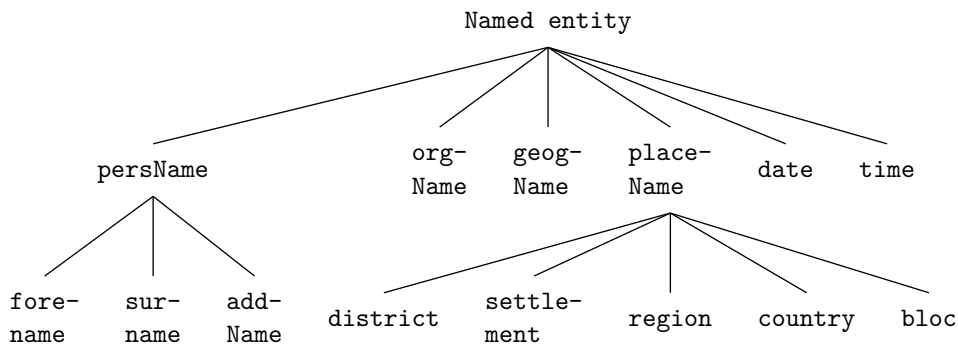
Definicja zakresu jednostek nazewniczych jako obiektów lingwistycznych jest kwestią sporną. Obok prototypowych nazw własnych, takich jak imiona, nazwiska czy nazwy geograficzne, występują określenia opisujące obiekty o tak zróżnicowanej naturze, jak instytucje, wytwory pracy ludzkiej, wydarzenia, jednostki miar i wag, momenty czasowe lub okresy itd. W projekcie NKJP, z uwagi na jego nowatorski charakter w odniesieniu do języka polskiego, zestaw tych kategorii został ograniczony do nazw osób, miejsc, instytucji i momentów czasowych.

[1] `http://www.itl.nist.gov/iad/mig/tests/ace/`.

W pracach badawczych w zakresie opisu i identyfikacji jednostek nazewniczych zaproponowano wiele taksonomii tych jednostek, od bardzo podstawowych zawierających kilka ogólnych klas, jak np. w kampaniach MUC (Chinchor 1997), do złożonych hierarchii wielopoziomowych, uwzględniających relacje, jak w wypadku programu ACE[2], ontologii GATE[3] czy prac Sekine'a i in. (2002) oraz Maurela (2008).

W NKJP przyjęliśmy taksonomię o średnim stopniu szczegółowości przedstawioną na rys. 9.2. Jest ona inspirowana wskazówkami formatu TEI P5[4], tzn. przejmuje niektóre elementy i ich semantykę zdefiniowaną w obrębie tego standardu dla jednostek nazewniczych. Poniżej opisany jest bardziej szczegółowo każdy z sześciu typów głównych oraz ośmiu podtypów.

Rysunek 9.2. Hierarchia typów polskich jednostek nazewniczych w NKJP



Nazwy osób

Do grupy tej, oznaczanej przez `persName`, należą nazwy indywidualnych osób i rodzin (w odróżnieniu od grup ludzkich klasyfikowanych jako organizacje):

1. rzeczywistych: *Zbigniew Hołdys, Bush Jr., Maria Skłodowska-Curie, Jan Paweł II, Kaczyńscy, Habsburgowie*;
2. fikcyjnych, legendarnych i pochodzących z wierzeń religijnych: *Bóg, Święty Jerzy, Archanioł Gabriel, Pan Wołodyjowski, Plastuś, Król Artur*.

Wyróżniamy w niej następujące podtypy, będące w istocie częściami składowymi pełnych nazw:

1. `forename` – imię, ewentualnie złożone, zdrobniałe lub w liczbie mnogiej: *Marcin, Kasia, Jean-Marie, Krzysztofowie*;

[2] `http://projects.ldc.upenn.edu/ace/annotation/`.

[3] `http://gate.ac.uk/`.

[4] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html`.

2. `surname` – nazwisko: *Kaczmarek, Kowalscy, Washington, Joice*;
3. `addName` – pseudonim, przydomek, dynastia, dodatkowy epitet: *Grot, Bez Ziemi, Lwie Serce, Groźny, Jagiellonowie, Waza*.

Części składowe, niemogące funkcjonować niezależnie jako nazwy własne, np. *van der, Junior* itp., nie mają odrębnego podtypu. Standard TEI definiuje dla takich składników odrębne elementy `nameLink` i `genName`, których nie wykorzystujemy, por. przykład (9.8).

Nazwy organizacji

Grupę tę, podobnie jak we wskazówkach TEI, oznaczamy znacznikiem `orgName` i nie dzielimy jej na podtypy. Zaliczamy do niej nazwy:

1. organizacji i instytucji międzynarodowych: *Parlament Europejski, NATO*;
2. organizacji i instytucji państwowych i publicznych: *Kancelaria Prezydenta RP, Polskie Koleje Państwowe, Zakład Ubezpieczeń Społecznych*;
3. instytucji i ośrodków kultury, nauki i szkolnictwa: *Uniwersytet Jagielloński, Teatr Bajka, Polska Akademia Nauk, Biblioteka Uniwersytecka*;
4. firm: *Hortex, Symantec Polska, Sony, Volvo, Boeing*;
5. stowarzyszeń, federacji i zespołów: *Towarzystwo Krzewienia Kultury Fizycznej, Zespół Pieśni i Tańca „Mazowsze", Maanam*;
6. zgrupowań wojskowych i paramilitarnych: *II Armia Wojska Polskiego, Batalion „Zośka", Czerwone Brygady*;
7. obiektów geograficznych i geopolitycznych lub wytworów pracy ludzkiej, gdy przybierają one znaczenie metonimiczne: *Polska, Unia Europejska, Ameryka, Gazeta Wyborcza*.

Nazwy geograficzne

Są to nazwy obiektów geograficznych mających cechy fizyczne wyróżniające je w terenie. Oznaczane są znacznikiem `geogName` i nie podlegają podziałowi na podtypy. Do grupy tej należą nazwy:

1. regionów historycznych i geograficznych, często niepokrywających się z podziałem administracyjnym: *Mazury, Podhale, Wielkopolska, Masyw Centralny, Wielka Nizina Węgierska, Bliski Wschód*;
2. wysp, mierzei, półwyspów, przylądków: *Tasmania, Mierzeja Helska, Kamczatka, Jawa, Przylądek Dobrej Nadziei*;
3. rzek, kanałów, źródeł, cieków wodnych: *Wisła, Kanał Augustowski*;
4. jezior, mórz, oceanów, prądów: *Bałtyk, Morze Kaspijskie, Morskie Oko, Lac Leman, Jezioro Wiktorii, Pacyfik, Ocean Lodowaty, Golfsztrom*;

5. ulic, dróg, autostrad, placów, skwerów, parków i innych obiektów miejskich: *Aleje Jerozolimskie, Plac Bankowy, Autostrada Słońca, Trasa Toruńska, Ucho Igielne, Łazienki*;
6. budynków, pomników i instalacji: *Pałac Kultury i Nauki, Wawel, Barbakan, Kaplica Zygmuntowska, Teatr Bajka, Statua Wolności*;
7. lasów, parków narodowych i krajobrazowych, rezerwatów, innych terenów zielonych: *Bory Tucholskie, Puszcza Białowieska, Tatrzański Park Narodowy*;
8. gór i szczytów: *Tatry, Mount Everest, Wieżyca*;
9. obiektów astronomicznych: *Słońce, Wielki Wóz, Morze Spokoju*.

Nazwy geopolityczne

Kategoria ta, oznaczana jako `placeName`, zawiera nazwy obiektów umotywowanych geopolitycznie, a więc wynikających z podziałów administracyjnych o różnym stopniu szczegółowości. Wyróżniamy wśród nich pięć podtypów:

1. `district` – jednostka podziału administracyjnego miasta lub innej osady, dokonanego według dowolnego kryterium, np. osiedle, dzielnica, parafia itp.: *Piaski, Gmina Bielany, Żoliborz, Za Żelazną Bramą, Powązki, Parafia Św. Zygmunta*;
2. `settlement` – pojedyncze miasto, wioska lub osada: *Katowice, Piaski, Nowa Słupia, Bordeaux*;
3. `region` – jednostka podziału administracyjnego większa niż miasto, ale mniejsza niż państwo, np. województwo, stan, prowincja: *województwo mazowieckie, gmina Pisz, powiat bieruńsko-lędziński, archidiecezja gnieźnieńska, Teksas, Bawaria*;
4. `country` – państwo lub kraj, kolonia, wspólnota: *Polska, Republika Czeska, RPA, Kongo, Zjednoczone Królestwo, Związek Radziecki*;
5. `bloc` – jednostka geopolityczna obejmująca dwa lub więcej państw: *Unia Europejska, NATO*.

Wyrażenia czasowe

Określenia związane z różnymi aspektami czasowymi stanowią obecnie przedmiot intensywnych badań, a w szczególności są objęte standardem TimeML[5], który zakłada bardzo szeroki zakres opisu tej problematyki. Wzięte są w nim pod uwagę tak złożone zjawiska, jak wydarzenia i ich „zaczepienie" na bezwzględnej osi czasu, wzajemna relacja i porządek wydarzeń, wyrażenia niedospecyfikowane (np. *w zeszłym tygodniu*), aspekty czasowników itd.

[5] `http://www.timeml.org`.

W projekcie NKJP ograniczyliśmy się do dużo prostszego zakresu anotacji zbliżonego do podzbioru propozycji TEI P5, tj. do oznaczeń:

1. `date` – dla dat kalendarzowych: *24 października 1945, grudzień 1981, 504 p.n.e., XXI wiek, rok 2000, piątego*;
2. `time` – dla określeń czasu w postaci godzin, minut i ewentualnie sekund: *12.25, pięć po dwunastej, ósma trzydzieści wieczorem,12.20, 9 sekund i 58 setnych*.

Wyrazy pochodne

Klasyfikacja niejako prostopadła do typów i podtypów zawiera wyrazy semantycznie związane z czterema pierwszymi typami i ich podtypami zaliczane do jednej z dwóch kategorii:

1. `relAdj` – przymiotniki relacyjne odnoszące się do nazw osób, organizacji i miejsc: *poznański, podwarszawski, proniemiecki, europejski, Chopinowski, ONZ-owski*;
2. `persDeriv` – nazwy mieszkańców miejsc geograficznych lub geopolitycznych, narodowości oraz członków organizacji: *poznaniak, Europejczyk, nie-Polak, Grek, żoliborzanin, Bawarczyk, AK-owiec*.

Jednostki te są istotnym elementem w przetwarzaniu tekstów, gdyż dają pełniejszy obraz wariantywności tych nazw. Na przykład w idealnym procesie ekstrakcji informacji dwa następujące wyrażenia powinny być rozpoznane jako równoznaczne:

(9.1) Muzeum Narodowe w Warszawie
warszawskie Muzeum Narodowe

Anotowaliśmy przymiotniki, które są natury relacyjnej, np. *francuski minister spraw zagranicznych*, a nie atrybutywnej, np. *francuski piesek*. Czasami to rozgraniczenie nie jest proste, jak pokazano w p. 9.4.4.

Czego nie anotowaliśmy

Kilka typów jednostek, które w innych projektach uznawane są za nazewnicze, nie było oznaczanych w NKJP, choć mogły podlegać anotacji ich części składowe. Do grupy tej należą:

1. tytuły i funkcje osób: *dyrektor* $[Adam\ Płocki]_{persName}$, *Prof.* $[A.\ Kamiński]_{persName}$;
2. nazwy zwierząt: *Freedom, Brooklyn i Diana Star* (tu: konie wyścigowe);
3. jednostki miar i wag (np. procenty, ceny, odległości): *pięć procent, 569,35 zł, 1,2 nm, kilowatogodzina*;

4. nazwy wytworów ludzkich (np. dzieła, produkty, pojazdy): *„Człowiek z marmuru”, Danonki, Volvo, Gazeta Wyborcza* – zauważmy, że w pewnych kontekstach niektóre z tych nazw mogą reprezentować instytucje i są wówczas oznaczone, por. przykład (9.36);
5. nazwy wydarzeń: *Rewolucja* $[Francuska]_{relAdj}$*, Powstanie styczniowe, Wielki Wybuch*;
6. daty opisowe: $[Boże]_{relAdj}$ *Narodzenie, Dzień Niepodległości*;
7. okresy – w przeciwieństwie do TEI, TimeML i MUC anotowaliśmy w nich jedynie odniesienie do konkretnych dat i godzin: *na przełomie* $[kwietnia]_{date}$ *i* $[maja]_{date}$*, od* $[dwunastej]_{time}$ *do* $[czternastej]_{time}$;
8. adresy: *ul.* $[[Wyspiańskiego]_{persName}]_{geogName}$ *32 m 12, www.msz.gov.pl*;
9. warianty stylistyczne nazw własnych: *Nowy Świat* (tzn. Ameryka), *Wenecja Północy, Ziemia Obiecana*;
10. liczba mnoga (jako generyczna) nazw o wspólnych członach: *Urzędy Wojewódzkie, Komisje Regionalne*;
11. określenia opisowe zachowujące zasadę jednostkowości wskazanego obiektu[6], ale zbyt dalekie od statusu nazw własnych: *kraje* $[śródziemnomorskie]_{relAdj}$*, episkopat*;
12. metafory: $[węgierski]_{relAdj}$ *Balcerowicz*;
13. rzeczowniki pochodne, które nie są nazwami mieszkańców, narodów czy członków organizacji: *hitleryzm, piłsudczyk*;
14. przymiotniki relacyjne mające niebezpośredni związek z nazwami osób, organizacji lub miejsc: *rzymskokatolicki, anglikański*;
15. przymiotniki pochodzące od wyrażeń czasu: *XII-wieczny, 12-godzinny*.

9.3.2. Reguły i strategie anotacyjne

Wszystkie jednostki nazewnicze występujące w korpusie i mieszczące się w zdefiniowanym powyżej zakresie były wydzielane w tekście, a następnie opisywane zgodnie z poniższymi regułami.

Zestaw atrybutów

Zidentyfikowanym jednostkom nazewniczym przypisywano atrybuty wymienione poniżej:

`@orth` – forma jednostki występująca w tekście, przypisywana do niej w sposób automatyczny, np. *Stanów Zjednoczonych*;

`@base` – forma podstawowa jednostki w sensie gramatycznym, np. *Stany Zjednoczone*;

[6] Por. obiekty typu „mentions” w programie ACE.

`@when` – data lub godzina znormalizowana zgodnie ze standardem ISO 8601 zalecanym przez W3C[7], w postaci `(-)yyyy-mm-dd` lub `hh:mm:ss(.s+)`, przy czym częściowe daty i godziny normalizowane były przez pominięcie elementu brakującego, np.:

(9.2) *dnia 21 września 1960 r.*: *@when=1960-09-21*
sierpień 80: *@when=1980-08*
56 p.n.e.: *@when=-0056*
piątego: *@when=---05*
ósma trzydzieści wieczorem: *@when=20:30:00*
godzinie jedenastej, 29 minut, 10 sekund i 5 setnych: *@when=11:29:10.05*

`@type` – jeden z sześciu typów głównych taksonomii przedstawionej na rys. 9.2;

`@persNameType` i `@placeNameType` – jeden z ośmiu podtypów w wypadku typu głównego `persName` lub `placeName`;

`@derivType` – typ wyrazu pochodnego, a więc `relAdj` lub `persDeriv`;

`@derivedFrom` – leksem, do którego odnosi się dany wyraz pochodny; może on być podstawą derywacji morfologicznej (np. *Poznań* dla *poznański*), ale czasem jest on umotywowany semantycznie a nie morfologicznie (np. *Stany Zjednoczone* dla *amerykański*); w dalszej części tej publikacji leksem ten nazywany jest, z pewną dozą nieścisłości, *bazą derywacyjną*;

`@cert` – poziom pewności anotacji, z możliwymi wartościami: `high`, `medium`, `low` i `unknown`; za pomocą tego atrybutu anotatorzy mogą wyrażać swoje wątpliwości wobec dokonanego oznaczenia danej jednostki;

`@certComment` – komentarz do poziomu pewności anotacji, np. *Brak pewności czy to imię czy nazwisko*.

Pełny zestaw atrybutów dla danej jednostki zależy od jej typu i kontekstu. Atrybuty `@orth` i `@type` są obowiązkowe dla wszystkich jednostek, choć nie zawsze ich ustalenie okazywało się proste (por. p. 9.4.3). Atrybut `@when` dotyczy wyłącznie typów `date` i `time`, natomiast atrybut `@base` jest obowiązkowy dla czterech pozostałych typów. Dla wyrażeń czasowych (np. *dnia 25 marca*) nie podawaliśmy formy podstawowej `@base` z uwagi na to, że mogą być one uznawane za wyrażenia nominalne, z formą podstawową w mianowniku (np. *dzień 25 marca*), albo też za wyrażenia przysłówkowe z formą podstawową identyczną formie tekstowej (np. *dnia 25 marca*). Atrybut `@placeNameType` jest obowiązkowy dla każdej nazwy typu `placeName`, inaczej natomiast ma się sprawa z atrybutem `@persNameType`. Jak już wspomniano, podtypy nazw osobowych opisują części składowe tych nazw, a nie podtypy nazwanych obiektów. Dlatego w pełnej nazwie osobowej, np. *Maria Skłodowska-Curie*, części składowe (np. *Maria*) otrzymywały atrybut

[7] `http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/#isoformats`.

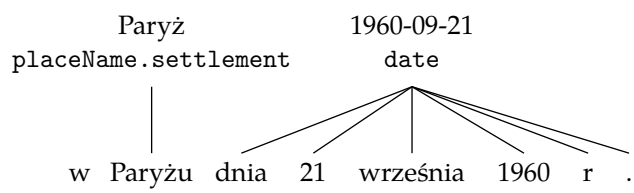
`@persNameType` (tu: `forename`), cała nazwa zaś opisywana była wyłącznie typem głównym `@type`. Atrybuty `@derivType` i `@derivedFrom` zarezerwowane są dla wyrazów pochodnych. I tak na przykład przymiotnik relacyjny *francuski* uzyskiwał atrybuty `@type=placeName`, `@placeNameType=country`, `derivType=relAdj` i `@derivedFrom=Francja`. Jak to opisano poniżej, wskazanie wartości tych atrybutów mogło czasem okazać się problematyczne.

W dalszej części tej publikacji przyjmujemy dwie równoważne konwencje notacyjne dla przykładów jednostek nazewniczych i ich atrybutów. Według pierwszej z nich, pokazanej w przykładach (9.3)–(9.4), anotowana jednostka jest okalana nawiasami kwadratowymi. Za nawiasem zamykającym na pozycji indeksu dolnego podawany jest ewentualny typ wyrazu pochodnego `relAdj` lub `persDeriv` oraz typ główny z ewentualnym podtypem. Na pozycji indeksu górnego pojawia się forma podstawowa lub znormalizowana jednostki oraz jej ewentualna baza derywacyjna. Druga konwencja, zilustrowana w przykładach (9.5)–(9.6), zakłada stworzenie drzewa anotacyjnego, w którym człony tekstu należące do jednej nazwy doczepiane są do wspólnego węzła etykietowanego najpierw formą podstawową lub znormalizowaną, następnie ewentualnym typem wyrazu pochodnego, typem głównym i ewentualnym podtypem, wreszcie ewentualną bazą derywacyjną.

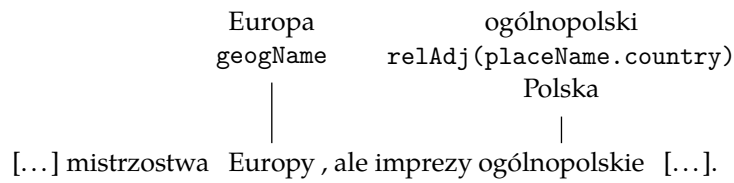
(9.3) w [Paryżu]$^{\text{Paryż}}_{\text{placeName.settlement}}$ [dnia 21 września 1960 r.]$^{\text{1960-09-21}}_{\text{date}}$

(9.4) [...] może nie od razu mistrzostwa [Europy]$^{\text{Europa}}_{\text{geogName}}$, ale [...] imprezy [ogólnopolskie]$^{\text{ogólnopolski; Polska}}_{\text{relAdj(placeName.country)}}$ rangi juniorskiej [...].

(9.5)



(9.6)



Jednostki zagnieżdżone

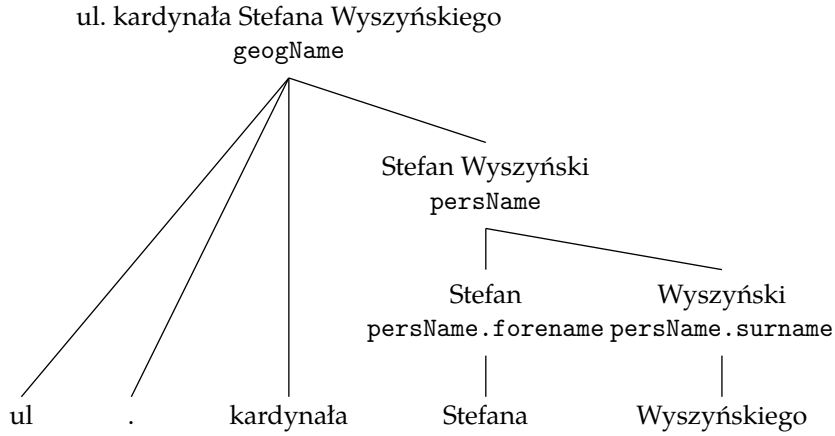
Jedną z nowatorskich cech przedstawionych tu reguł anotacji jest to, że anotowane były nie tylko jednostki nazewnicze o maksymalnej długości, ale również

wszystkie nazwy będące ich częściami składowymi, jak pokazują przykłady (9.7)–(9.9). Oznacza to w szczególności, że w obrębie jednostek niepodlegających anotacji mogą być oznaczane ich części składowe, jak w przykładach (9.10)–(9.11). Wydaje się, że tak opisany korpus ma co najmniej trzy zalety: a) cechuje go stosunkowo duża gęstość i różnorodność oznaczonych jednostek, b) ułatwia analizę nawiązań, jak i wzajemnych zależności opisywanych jednostek, c) może ułatwić ujednoznacznianie typów jednostek nazewniczych. Szczegółowa problematyka związana z anotacją nazw zagnieżdżonych jest omówiona w p. 9.4.7.

(9.7) $[[\text{Irlandzka}]^{\substack{\text{irlandzki; Irlandia}\\\text{relAdj(placeName.country)}}}$ Armia Republikańska$]^{\substack{\text{Irlandzka Armia Republikańska}\\\text{orgName}}}$

(9.8) $[[\text{Izaaka}]^{\substack{\text{Izaak}\\\text{persName.forename}}}$ van der $[\text{Blocke}]^{\substack{\text{Blocke}\\\text{persName.surname}}}]^{\substack{\text{Izaak van der Blocke}\\\text{persName}}}$

(9.9)



(9.10) Rewolucja $[\text{Francuska}]^{\substack{\text{francuski; Francja}\\\text{relAdj(placeName.country)}}}$

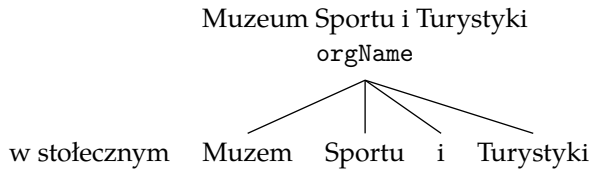
(9.11) kraje $[\text{śródziemnomorskie}]^{\substack{\text{śródziemnomorski; Morze Śródziemne}\\\text{relAdj(geogName)}}}$

Struktury spójnikowe z nakładającymi się elementami

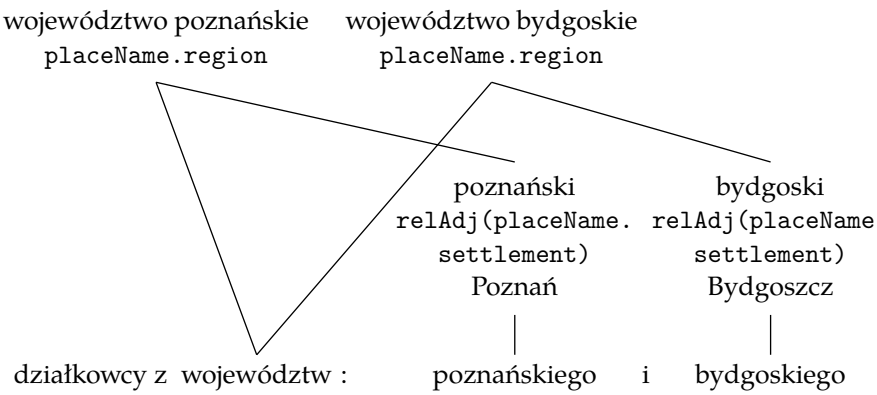
Inna oryginalna reguła anotacji w NKJP dotyczy oznaczania jednostek nazewniczych pojawiających się w strukturach spójnikowych. Jeśli spójnik jest integralną częścią nazwy, pojawia się on w drzewie anotacyjnym, jak w przykładzie (9.12). Jeśli zaś chodzi o dwa różne obiekty nazwane – zob. przykład (9.13) – to każdy z nich jest anotowany z osobna. Jest to szczególnie istotne, gdy tak skoordynowane nazwy dzielą wspólny człon. Pojawiają się wówczas dwa problemy: częściowe nakładanie się nazw oraz ich nieciągłość, których reprezentacja nie jest możliwa za pomocą nawiasowania, a jedynie przez wskaźniki z poziomu drzew anotacyjnych.

Zauważmy, że dla odzwierciedlenia takich zjawisk szczególnie przydatna jest anotacja typu *stand-off* (tekst źródłowy nie jest modyfikowany, anotacje zaś zawarte są w zewnętrznych plikach zawierających wskaźniki do pliku źródłowego).

(9.12)



(9.13)



Formy podstawowe

Z uwagi na warianty graficzne i gramatyczne nazw, niekiedy dana nazwa może mieć kilka form podstawowych. W NKJP założyliśmy, że ustalanie formy podstawowej odbywa się wyłącznie na płaszczyźnie fleksyjnej, tzn. sprowadza się w większości wypadków do podania mianownika liczby pojedynczej, a wariantywność innego rodzaju jest zachowywana. Oznacza to przyjęcie następujących zasad ustalania atrybutu `@base`:

1. nierozwijanie skrótów – akronimy i skróty nie są rozwijane do pełnej formy (zob. też przykład (9.9)):

   (9.14) $[\text{Ruch Społeczny } [\text{AWS}]^{\textbf{AWS}}_{\text{orgName}}]^{\text{Ruch Społeczny } \textbf{AWS}}_{\text{orgName}}$

2. nierozwijanie elips – formy eliptyczne nie są przekształcane na pełne nazwy:

   (9.15) Martwi nas dwulicowość radnych $[\text{Platformy}]^{\textbf{Platforma}}_{\text{orgName}}$.

3. nienormalizowanie zdrobnień – zdrobniałe imiona, nazwiska, przydomki itd. nie są sprowadzane do form nienacechowanych:

   (9.16) $[\text{Franusia}]^{\textbf{Franuś}}_{\text{persName.forename}}$

4. włączanie cudzysłowów – jeśli nazwa pisana jest w cudzysłowie, włączany on jest do nazwy i jej formy podstawowej:

   (9.17) polityczny reprezentant [„Solidarności”]$^{\textbf{„Solidarność”}}_{\text{orgName}}$

5. pisownia wielką lub małą literą – formy podstawowe przymiotników relacyjnych zawsze pisane są małą literą (i w liczbie pojedynczej rodzaju męskiego):

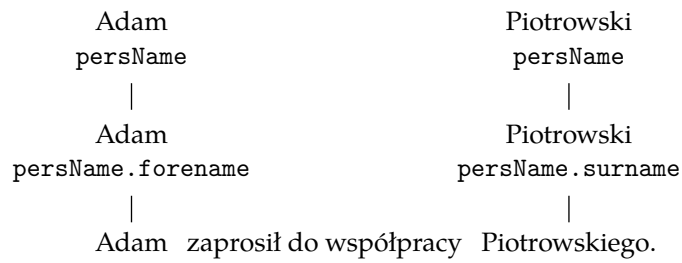
   (9.18) [[Warszawskie]$^{\textbf{warszawski}; \text{ Warszawa}}_{\text{relAdj(placeName.settlement)}}$
   Zakłady Optyczne]$^{\text{Warszawskie Zakłady Optyczne}}_{\text{orgName}}$

Strategie dotyczące nazw osobowych

Nazwy osobowe wyróżniają się spośród innych jednostek nazewnicznych kilkoma cechami szczególnymi. Jak już wspominaliśmy, ich podtypy opisują w istocie ich części składowe. Dodatkowo ich morfologia (np. odmiana przez liczbę) jest bogatsza niż dla innych nazw i częściej pojawiają się one w grupach spójnikowych. Dlatego przyjęto zestaw specyficznych reguł dotyczących tych nazw:

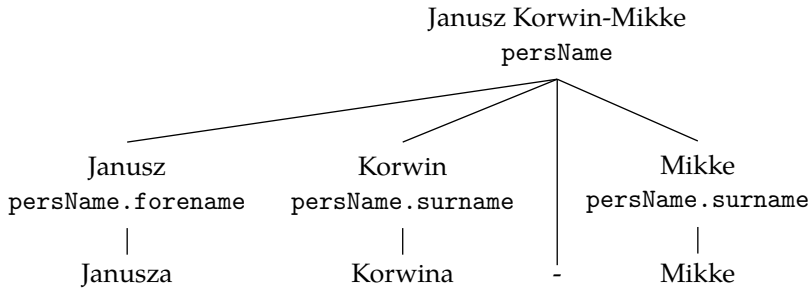
1. Anotacja pojedynczych nazw osobowych.
   Dla nazw osobowych tworzone są zawsze co najmniej dwa węzły: najwyższy poziom reprezentuje całą nazwę, a poziom pośredni jej części składowe; dzieje się tak również wtedy, gdy w tekście pojawiają się pojedyncze imiona, nazwiska lub przydomki:

   (9.19)

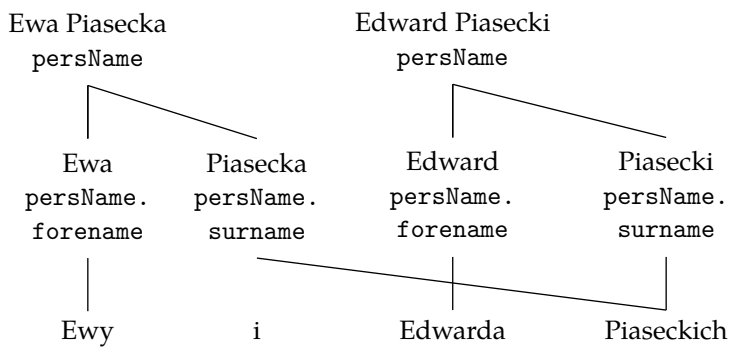
2. Anotacja nazwisk złożonych.
   Zakładamy w tym wypadku, że najpierw każda część składowa nazwiska jest anotowana osobno jako `persName.surname`, następnie są one bezpośrednio podłączane, wraz z dywizem, do nazwy głównej; nie jest zatem tworzony pośredni węzeł zawierający nazwisko złożone:

(9.20)



3. Cechy fleksyjne formy podstawowej nazwiska.
   Nazwisko jest zawsze uznawane za rzeczownik. Zatem jego forma podstawowa w mianowniku zachowuje rodzaj (męski lub żeński) danego wystąpienia, również wówczas, gdy nazwisko to ma pochodzenie przymiotnikowe (np. formą podstawową dla *Piaseckiej* jest *Piasecka*, a nie *Piasecki*). Zgodnie z przykładem (9.13), osoby o tym samym nazwisku opisane przy pomocy spójnika są anotowane osobno, jak to pokazuje przykład (9.21). Natomiast gdy nazwisko występuje w liczbie mnogiej bez wzmianki o imionach, za formę podstawową przyjmujemy mianownik liczby mnogiej – zob. przykład (9.22).

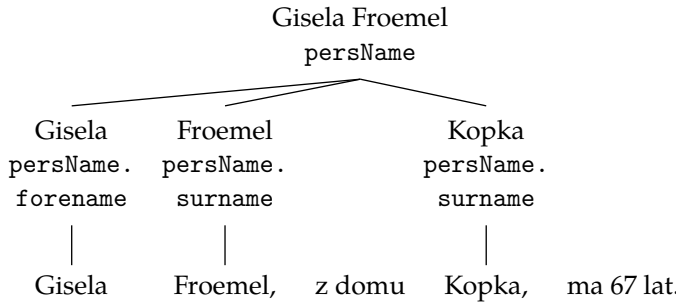
(9.21)



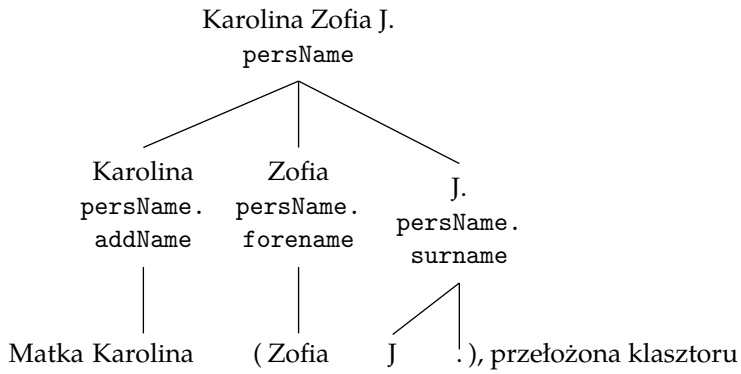
(9.22) Najmłodsi [[Kowalscy]$^{\textbf{Kowalscy}}_{\text{persName.surname}}$]$^{\textbf{Kowalscy}}_{\text{persName}}$ dostali jeszcze od rodziców po małej ciupadze.

4. Szczególne przypadki nazw osobowych.
   Zdarzają się one w odniesieniu do nazwisk panieńskich, herbów oraz przybranych imion. Oto przykłady anotacji zaproponowanych dla tych przypadków:

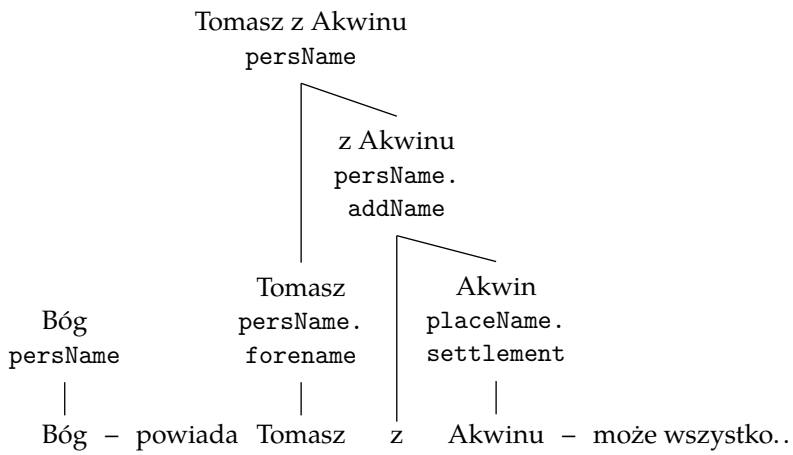
(9.23)



(9.24)



(9.25)



5. Wyjątkowość form podstawowych
   W szczególnych przypadkach przydomki języka potocznego mogą mieć formę podstawową w wołaczu:

   (9.26) $[[\text{Kwachu}]^{\textbf{Kwachu}}_{\text{persName.addName}}]^{\textbf{Kwachu}}_{\text{persName}}$ został [...].

9.4. Własności jednostek nazewniczych w kontekście NKJP

Anotacja nazw własnych na zasadach opisanych powyżej jest z formalnego punktu widzenia problemem klasyfikacji. Każda pojedyncza lub złożona jednostka korpusu, która została uznana za nazewniczą, była klasyfikowana, zgodnie z przyjętymi strategiami, jako należąca do dokładnie jednego z ustalonych typów i/lub podtypów. Następnie należało dla niej wybrać jedną wartość dla każdego atrybutu. Jednostki nazewnicze są jednak obiektami lingwistycznymi o rozmytych granicach i często trudnych do uszeregowania typach i własnościach. Poniżej opisujemy ciekawe zjawiska, na które natrafiliśmy w trakcie anotacji korpusu, utrudniające stosowanie opisanych wcześniej założeń metodologicznych.

9.4.1. Metonimia

Zjawiskiem, które sprawiało wiele trudności na różnych etapach anotowania, m.in. przy określaniu typu i/lub podtypu nazwy czy też przy wskazywaniu bazy derywacyjnej anotowanych przymiotników, jest metonimia. Ze strukturalistycznego punktu widzenia (Polański 1993):

> Metonimia jest odchyleniem modyfikującym strukturę ciągu syntagmatycznego, którego istotą jest redukcja: powstaje ona w wyniku niewypełnienia pozycji syntaktycznej przez jakieś wyrażenie semantycznie zgodne (kompatybilne) z wyrażeniem pozycję tę otwierającym i przesunięcia do niej innego wyrażenia, które z wyrażeniem brakującym wchodziłoby w tym samym ciągu w relację syntaktyczną (zajmowałoby przy nim pozycję).

W ujęciu kognitywnym metonimia – tak jak metafora – jest jednym z podstawowych sposobów posługiwania się językiem naturalnym. Z metonimią mamy do czynienia w sytuacji, kiedy – jak to formułują w prosty sposób Lakoff i Johnson (2010) – mówimy o pewnej rzeczy, używając innej rzeczy, która jest z nią związana. Metonimia „pełni głównie funkcję desygnacyjną, co powala nam używać pewnego pojęcia tak, aby zastąpiło inne" (Lakoff i Johnson 2010: 68).

Najczęściej spotykanymi w projekcie NKJP typami metonimii były: część za całość, (9.27), instytucja za osoby odpowiedzialne, (9.28), miejsce za instytucję, (9.29), miejsce za wydarzenie, (9.30).

(9.27) nowa twarz [telewizji]$^{\text{telewizja}}_{\textbf{orgName}}$

(9.28) [Sejm]$^{\text{Sejm}}_{\textbf{orgName}}$ przyjął ustawę.

(9.29) Choćby wtedy, gdy [Moskwa]$^{\text{Moskwa}}_{\textbf{orgName}}$ kazała żołnierzom rozpędzać saperskimi łopatkami studentów [...].

(9.30) Katyń[8] łączy Polaków.

Zgodnie ze strategią przyjętą w projekcie NKJP nazwy własne użyte metonimicznie oznaczane były typami zgodnymi z interpretacją właściwą dla danego użycia. Najczęściej stosowany był w takich wypadkach typ `orgName`, gdyż konteksty wskazywały na ludzi związanych z desygnatem danej nazwy (np. mieszkańców jakiegoś kraju, miasta lub pracowników jakiejś instytucji). I tak w przykładzie (9.31) jednostki *Niemcy* i *Kazachstan* były opisywane nie jako nazwy państw, lecz jako organizacje, ponieważ – co oczywiste – w meczu nie biorą udziału dwa kraje ani wszyscy ich mieszkańcy, tylko dwie drużyny piłkarskie złożone z wybranych przedstawicieli obu państw (jest to przykład niejako podwójnej metonimii: miejsce za mieszkańców i część za całość). Zdanie (9.32) jest również metonimiczne, przy czym formę *Niemcy* opisano jako derywację osobową, tj. nazwę mieszkańców/obywateli utworzoną od nazwy państwa *Niemcy*. W obu wypadkach pomocna we wskazaniu właściwego typu i podtypu nazwy była forma rodzajowa czasownika *pokonać* w czasie przeszłym: rodzaj męskoosobowy wskazuje na nazwę mieszkańców/obywateli, a rodzaj niemęskoosobowy na nazwę kraju.

(9.31) [Niemcy]$^{\text{Niemcy}}_{\textbf{orgName}}$ pokonały [Kazachstan]$^{\text{Kazachstan}}_{\textbf{orgName}}$ 3-0.

(9.32) [Niemcy]$^{\text{Niemiec; Niemcy}}_{\textbf{persDeriv(placeName.country)}}$
pokonali [Szwedów]$^{\text{Szwed; Szwecja}}_{\textbf{persDeriv(placeName.country)}}$ 3:2.

W korpusie NKJP wiele było kontekstów, w których nazwy miast, zwłaszcza stolic państw, występowały w znaczeniu rządów, przywódców, osób odpowiedzialnych za politykę danego kraju. Obrazuje to przykład (9.33). Z kolei nazwa *Bruksela* była używana w znaczeniu instytucji europejskich, które mają swoje siedziby w tym mieście, czyli Parlamentu Europejskiego lub Komisji Europejskiej, por. (9.34).

(9.33) Czekamy na reakcję [Wilna]$^{\text{Wilno}}_{\textbf{orgName}}$ i [Kijowa]$^{\text{Kijów}}_{\textbf{orgName}}$.

(9.34) Wniosek władz toruńskich [...] trafił już do [Brukseli]$^{\text{Bruksela}}_{\textbf{orgName}}$.

Metonimia sprawiła, że opisowi w ramach projektu NKJP podlegały nawet te nazwy, których w użyciach niemetonimicznych z zasady nie anotowaliśmy. Dotyczy to m.in. tytułów gazet lub informacyjnych programów telewizyjnych. Na przykład, o ile w wypadku (9.35) mamy do czynienia z tytułem gazety, którego

[8] Katyń jako wydarzenie nie jest anotowany.

nie anotujemy, o tyle w zdaniu (9.36) tytuł dziennika został użyty metonimicznie i oznacza instytucję, redakcję lub osoby z nią związane – dziennikarzy.

(9.35) W sobotnim „Głosie”[9] pisaliśmy o […]

(9.36) Powiedział wczoraj [„Dziennikowi”]$^{\text{„Dziennik”}}_{\textbf{orgName}}$, że […].

Bywa i odwrotnie, kiedy to – zgodnie z przyjętymi w projekcie regułami – nazwa zazwyczaj anotowana, użyta metonimicznie, nie jest opisywana. Ilustrują to przykłady (9.37) i (9.38), w których nazwisko oznacza nie autora, lecz jego nowe dzieło. Jest to kolejny przykład metonimii (Lakoff i Johnson 2010: 67).

(9.37) Poczytajmy trochę nowego Ripaldę, żeby się oświecić.

(9.38) Z optymistycznych książek – nowa Grochola się pokazała.

Jedną z najbardziej skomplikowanych kwestii była anotacja jednostek nazewniczych związanych z Europą i Unią Europejską. Przyczyną tej trudności były użycia metonimiczne leksemu *Europa* oraz przymiotnika *europejski*. Jednostki te niezmiernie rzadko oznaczały kontynent, jak w przykładzie (9.39). Najczęściej odnosiły się one do mieszkańców lub obywateli Europy jako kontynentu w sensie geograficznym (9.40) lub jako organizacji. W tym ostatnim przypadku nazwa *Europa* była często synonimem *Unii Europejskiej*, interpretowanej jako blok państw lub jako instytucja. Obrazują to przykłady (9.41) i (9.42). Ze względu na przyjęte standardy i hierarchię typów nazw w każdym z omówionych przypadków jednostka *Europa* była anotowana jako `orgName`.

(9.39) […] odnalazłem przylądek św. Wincentego – miejsce, gdzie kończy się [Europa]$^{\text{Europa}}_{\textbf{geogName}}$ […].

(9.40) [Europa]$^{\text{Europa}}_{\textbf{orgName}}$ i Ameryka miały inne sprawy na głowie. Trwała światowa wojna.

(9.41) Tym się różnimy od [Europy]$^{\text{Europa}}_{\textbf{orgName}}$ i tę polską odrębność [Europa]$^{\text{Europa}}_{\textbf{orgName}}$ powinna poznać.

(9.42) Raport Komisji Europejskiej pt. „Edukacja dla Europy” postuluje: „[Europa]$^{\text{Europa}}_{\textbf{orgName}}$ powinna popierać ideę, iż wydatki na edukację […] powinny osiągnąć w ciągu następnych 10 lat przeciętnie 8%”.

Podobnych trudności nastręcza anotacja przymiotnika *europejski*, którego baza derywacyjna zmienia się w zależności od kontekstu. Jest to typową własnością przymiotników, które nabierają znaczenia dopiero w kontekście, tj. w połączeniu z rzeczownikiem. W zdaniu (9.43) przymiotnik *europejski* odnosi się do instytucji Parlamentu Europejskiego – możliwe jest uzupełnienie: „kolejne wybory do

[9] „Głos” nie jest anotowany, ponieważ jest nazwą wytworu ludzkiego.

Parlamentu Europejskiego". Natomiast w użyciach (9.44) i (9.45) trudno wskazać jeden właściwy opis bazy derywacyjnej – wyrażenie *standardy europejskie* może znaczyć 'takie jak w Europie' (rozumianej geograficznie jako kontynent), 'takie jak w krajach Unii Europejskiej' (interpretowanej także jako blok państw) lub 'zgodne z ustaleniami Unii Europejskiej' (w znaczeniu instytucji). Z kolei połączenie *europejski poziom* może być opisane – podobnie jak przykład wcześniejszy – jako 'taki jak w Europie' (rozumianej geograficznie jako kontynent lub geopolitycznie jako blok) albo 'taki jak w Unii Europejskiej', przy czym w tym wypadku jednostka *Unia Europejska* ma raczej znaczenie geopolityczne, a nie instytucjonalne.

(9.43) Wkrótce odbędą się kolejne [europejskie]$^{\text{europejski; }\textbf{Parlament Europejski}}_{\text{relAdj(}\textbf{orgName}\text{)}}$ wybory.

(9.44) Sklep ma być przestronny, nowoczesny, spełniający standardy [europejskie]$^{\text{europejski; }\textbf{Unia Europejska}}_{\text{relAdj(}\textbf{orgName}\text{)}}$ i przyjazny klientowi.

(9.45) Spała, Wałcz czy Cetniewo są od dawna ośrodkami na najwyższym [europejskim]$^{\text{europejski; }\textbf{Unia Europejska}}_{\text{relAdj(}\textbf{placeName.bloc}\text{)}}$ poziomie.

9.4.2. Elipsa

Elipsa, tj. pominięcie pewnego elementu wyrażenia syntaktycznego, często występuje w NKJP. W korpusie znajdujemy wiele przykładów jednowyrazowych wariantów eliptycznych nazw własnych, zwłaszcza partii politycznych (9.46) i instytucji państwowych (9.47), które od wyrazów pospolitych pozwala odróżnić przede wszystkim kontekst oraz – gdy kontekst nie jest wystarczająco czytelny – użycie wielkiej litery:

(9.46) Platforma (Obywatelska), Liga (Polskich Rodzin), Sojusz (Lewicy Demokratycznej)

(9.47) Rada (Polityki Pieniężnej), Komisja (Europejska / Państwowa Komisja Wyborcza), Izba (Krajowa Izba Kontroli), Uniwersytet (Wrocławski)

Nazwy partii i organizacji oraz ich warianty eliptyczne bywają oczywiście używane metaforycznie i metonimicznie – w znaczeniu osób związanych z daną organizacją, odpowiedzialnych za jej funkcjonowanie. Ten rodzaj metonimii nie prowadzi jednak do zmiany typu. Obrazują to zjawisko następujące przykłady:

(9.48) [Platforma]$^{\text{Platforma}}_{\textbf{orgName}}$ konsekwentnie mówi o swoim programie naprawy sytuacji w Polsce.

(9.49) [Komisja]$^{\text{Komisja}}_{\textbf{orgName}}$ wytknęła jeszcze jedną nieścisłość – w oświadczeniu majątkowym wicemarszałka Antoniego Pietkiewicza za rok 2002.

Związki metonimii z wariantami eliptycznymi można dostrzec także w nazwach klubów sportowych zawierających inne nazwy własne, np.: pełna nazwa *Wisła Kraków* – w skrócie *Wisła*. W projekcie NKJP całościowe nazwy tego typu anotowaliśmy jako organizacje z zagnieżdżeniem nazwy geograficznej oraz nazwy miasta – zob. (9.50) – lub jedynie nazwy miasta – zob. (9.51). Z uwagi na trudności interpretacyjne nie zostały określone kryteria pozwalające dokonać jednoznacznego wyboru między tymi dwiema strategiami.

(9.50) $[[\text{Wisła}]^{\text{Wisła}}_{\textbf{geogName}}\ [\text{Kraków}]^{\text{Kraków}}_{\textbf{placeName.settlement}}]^{\text{Wisła Kraków}}_{\textbf{orgName}}$

(9.51) $[\text{Wisła}\ [\text{Kraków}]^{\text{Kraków}}_{\textbf{placeName.settlement}}]^{\text{Wisła Kraków}}_{\textbf{orgName}}$

W zdaniach, w których występowała nazwa klubu z elipsą miasta, nie zagnieżdżaliśmy nazwy geograficznej – od razu opisywaliśmy interesującą nas jednostkę jako organizację. Przykład (9.52) jest przypadkiem granicznym, który można interpretować albo jako użycie metonimiczne, albo jako elipsę. Innymi słowy, trudno jest stwierdzić, czy *Wisła* stanowi nazwę organizacji, ponieważ chodzi o klub leżący nad Wisłą (metonimia), czy też z tego powodu, iż jest elipsą nazwy organizacji *Wisła Kraków*.

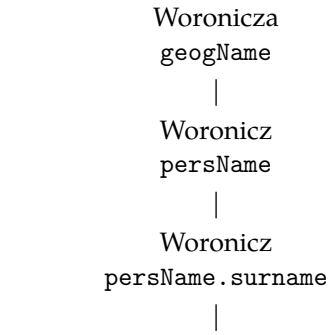
(9.52) Jednak w ostatnim meczu w Łodzi $[\text{Wisła}]^{\text{Wisła}}_{\textbf{orgName}}$ zagrała dobrze i mam nadzieję, że tę formę podtrzyma.

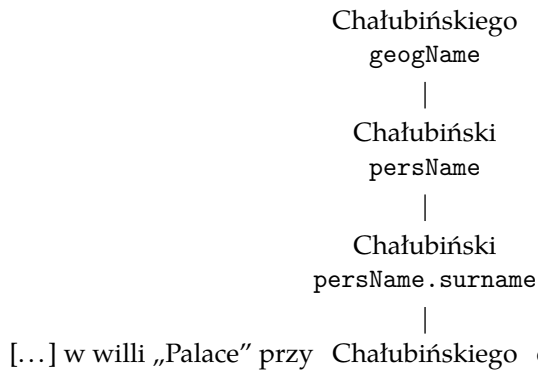
To samo dotyczyło nazw klubów, które w wersji eliptycznej zawierały jedynie nazwę miasta. Nazwę całościową anotowaliśmy z zagnieżdżeniem (9.53), nazwę z elipsą zaś bez zagnieżdżenia (9.54).

(9.53) W wyniku rekomendacji trenera Fernando Castro Deco przeniósł się do $[\text{FC}\ [\text{Porto}]^{\text{Porto}}_{\textbf{placeName.settlement}}]^{\text{FC Porto}}_{\textbf{orgName}}$.

(9.54) W rewanżowym meczu kwalifikacyjnym 3. rundy Ligi Mistrzów $[\text{Barcelona}]^{\text{Barcelona}}_{\textbf{orgName}}$ pokonała $[\text{Wisłę}]^{\text{Wisła}}_{\textbf{orgName}}$ 1-0.

Innym przykładem elipsy są użycia nazw ulic zawierających w swojej strukturze nazwisko z pominięciem składnika *ulica*, *aleja*. Prawidłowa anotacja jednostek *Woronicza* i *Chałubińskiego* w przykładach (9.55) i (9.56) powinna zawierać lemat *Woronicza* i *Chałubińskiego* (co wynika ze związku składniowego i wymagania dopełniacza), a dopiero zagnieżdżone w nich nazwy osobowe otrzymują lematy w formie mianownikowej: *Woronicz* i *Chałubiński*. Warto przy tym zaznaczyć, nawiązując do omawianego w p. 9.4.1 problemu metonimii, że zdanie (9.55) zawiera metonimiczne użycie nazwy ulicy. W przywołanym kontekście jednostka ta oznacza de facto (bądź może oznaczać – na podstawie kontekstu kulturowego, wiedzy o świecie) Telewizję Publiczną, instytucję mieszczącą się przy ul. Woronicza 17.

(9.55)


(9.56)


9.4.3. Niejednoznaczność typów i podtypów nazw

Zastosowanie w praktyce opracowanej na potrzeby projektu taksonomii typów i podtypów jednostek nazewniczych było jednym z podstawowych zadań anotatorów. Sprawiało przy tym niejednokrotnie spore trudności.

O kłopotach z oznaczaniem typów i podtypów nazw w wypadku jednostek geopolitycznych o spornym statusie, jak np. Kosowo, piszemy w p. 9.4.9. Wybór typu i/lub podtypu anotowanej nazwy utrudniało ponadto zjawisko metonimii, o którym wspominaliśmy w p. 9.4.1. Problem ten pojawiał się zwłaszcza w odniesieniu do jednostki *Europa* oznaczanej jako nazwa geograficzna, geopolityczna lub nazwa organizacji. Podobnie jest w wypadku nazwy *Anglia*, która czasami występuje w znaczeniu jednostki geopolitycznej (regionu, części składowej Wielkiej Brytanii), jak w przykładzie (9.57), ale najczęściej używana jest metonimicznie jako synonim tego państwa – por. (9.58).

(9.57) Dobre występy w lidze szkockiej, dobry występ z $[\text{Anglią}]^{\text{Anglia}}_{\text{placeName.}\mathbf{region}}$ w „kotle" Old Trafford [...].

(9.58) Między modelem Unii, do którego dążą Niemcy, Włochy i kraje Beneluksu, a tym, do którego pragną ją zredukować [Anglia]$^{\text{Anglia}}_{\text{placeName.}\mathbf{country}}$, Szwecja i Dania, istnieje prawdziwa przepaść.

Osobną kwestią jest określanie podtypów w ramach nazw osób. Na ogół wskazanie imienia (np. *Margaret*, *Adam*, *Krzyś*), nazwiska (np. *Thatcher*, *Małysz*, *Morawski*) lub przydomku (np. *Groźny*, *Lwie Serce*, *Madonna*) nie sprawiało trudności. Jednak w niektórych obcych nazwach osobowych, zwłaszcza wielosegmentowych, np. (9.59), rozpoznanie, który segment nazwy odpowiada imieniu (imionom), który nazwisku (nazwiskom), bywało kłopotliwe.

(9.59) Ahmad Hussein Khudayir as-Samarrai

Innym problemem był opis nazwy typu (9.60) i związany z nim kontekst kulturowy, a także stopień włączania do analizy tego, co wynika z wiedzy o świecie. Nazwa ta składa się z imienia i nazwiska, ale jednocześnie jej całość jest pseudonimem pisarza i publicysty – z tego względu mogła być określona jako przydomek (`addName`).

(9.60) Bolesław Prus

Dużo większe wątpliwości budziła anotacja jednostek typu *Zachód* i *Wschód*. Nazwy te można interpretować i opisywać – w zależności od kontekstu – na kilka sposobów. W sensie geopolitycznym jest to grupa państw (chodzi przede wszystkim o rozwinięte kraje Europy Zachodniej oraz Stany Zjednoczone). Ponieważ określenie, o jakie państwa dokładnie chodzi, jest problematyczne, a dodatkowo kraje te nie należą do jednej struktury politycznej, nie możemy się do nich odwoływać formalnie jako do bloku. Opisujemy zatem *Zachód* jako nazwę geograficzną określającą pewną część kontynentu europejskiego i Ameryki Północnej (9.61). Kiedy natomiast jednostki *Zachód* i *Wschód* użyte są metonimicznie, klasyfikowane są one jako skupiska ludzkie, czyli organizacje (9.62). W takich wypadkach kontekst wskazuje nie tyle na przestrzeń geograficzną czy grupę państw, ile na ludzi oraz wykształcone na obszarach zwanych *Zachodem* i *Wschodem* systemy wartości – w wymiarze kulturowym, gospodarczym i cywilizacyjnym.

(9.61) Ja nie byłem nigdy na [Zachodzie]$^{\text{Zachód}}_{\mathbf{geogName}}$.

(9.62) [Zachód]$^{\text{Zachód}}_{\mathbf{orgName}}$ mylnie pojmuje religię hinduską jako politeizm.

Podobne zasady dotyczą wyboru typu dla nazw kontynentów. Ponieważ niektóre państwa, jak Turcja czy Rosja, leżą na granicy dwóch kontynentów,

jednostki *Europa, Azja, Afryka* itd. nie mogą być oznaczane jako bloki państw, ale jako nazwy geograficzne (lub nazwy zbiorowości ludzkich)[10]:

(9.63) Najwięcej uchodźców, jest w [subsaharyjskiej]$^{\text{subsaharyjski; Sahara}}_{\text{rejAdj(geogName)}}$ [Afryce]$^{\text{Afryka}}_{\textbf{geogName}}$.

9.4.4. Niepewny status i pochodzenie przymiotników relacyjnych

Jak wspomniano w p. 9.3.1, w projekcie NKJP anotowane były przymiotniki, które pochodziły od nazw własnych zakwalifikowanych do jednej z czterech klas: osób, organizacji, miejsc geograficznych lub geopolitycznych, oraz miały charakter relacyjny, np. *francuski pułkownik*. Anotacji nie podlegały natomiast przymiotniki natury atrybutywnej w wyrażeniach typu *francuski piesek*. Jednak ustalenie granicy pomiędzy tymi dwoma grupami nie zawsze jest łatwe. Niekiedy nie sposób jednoznacznie określić, czy przymiotnik ma jeszcze charakter relacyjny, czy już atrybutywny. Jako przykład takich jednostek mogą służyć nazwy przepisów kulinarnych odnoszące się do kraju (regionu) ich rzekomego lub faktycznego pochodzenia, np.:

(9.64) Dla skądinąd sympatycznego stewarda istniał tylko jeden rodzaj kawy – po [turecku]$^{\text{turecki, }\textbf{Turcja}?}_{\textbf{relAdj(placeName.country)}}$.

W niektórych przypadkach ustalenie natury przymiotnika wymaga szerszego kontekstu. Przymiotnik w wyrażeniu *angielska herbata* skłonni jesteśmy interpretować jako atrybutywny – por. (9.65), tymczasem ścisły związek znaczeniowy łączący go z bazą morfologiczą może czasem zostać zachowany – por. (9.66).

(9.65) Dobra angielska herbata nie rośnie w Anglii, podobnie jak egzotyczne przyprawy.

(9.66) [...] popijali [angielską]$^{\text{angielski, Anglia}}_{\textbf{relAdj(placeName.region)}}$ herbatę z paczek, które dostawałem od żony z Anglii [...].

Inny problem dotyczy mnogości możliwych baz derywacyjnych. W wypadku niektórych wyrazów utworzonych od nazw miejscowych wręcz niemożliwe stawało się ustalenie jednostki będącej podstawą derywatu. Jako przykład niech posłuży przymiotnik *ostrowski*. Według słownika pod redakcją Kubiak-Sokół i Łazińskiego 2007 w Polsce występuje kilka miejscowości, dla których jest to przymiotnik relacyjny: *Ostrów Wielkopolski*, *Ostrów Lubelski*, *Ostrów Mazowiecka*, *Ostrów*, *Ostrowo*, *Ostrowo Kościelne*, *Ostrowo Mogileńskie*, *Ostrowsko*, *Ostrowy nad Okszą* i *Ostrowy Tuszowskie*. Kontekst (9.67) mógł pozwalać na ograniczenie tego

[10] Wyjątek stanowi nazwa *Europa*, której w uproszczeniu używa się w odniesieniu do Unii Europejskiej.

zbioru np. do miast powiatowych (tu: *Ostrów Mazowiecka* lub *Ostrów Wielkopolski*), tym niemniej jednoznaczne ustalenie odpowiedniej bazy mogło się okazać niemożliwe. Przypomnijmy, że kontekst umożliwiający ewentualne ujednoznacznienie jest w NKJP ograniczony z uwagi na losowy wybór próbek oraz ich rozmiar nie przekraczający pojedynczego akapitu.

(9.67) Sąsiedni powiat [ostrowski]$^{\text{ostrowski, }\textbf{Ostrów}?}_{\text{relAdj(placeName.settlement)}}$ w tym temacie jest przodującym w całym kraju, a więc pieniądze można zdobyć, trzeba tylko chcieć.

Tego typu jednostek jest oczywiście więcej: *wadowicki* może pochodzić od *Wadowice* lub *Wadowice Górne*, *gorzowski* od wsi *Gorzów*, *Gorzowa Wielkopolskiego* lub *Gorzowa Śląskiego* itd. W takich wypadkach anotacja oznaczała czasami konieczność rezygnacji z uszczegółowiania (np. *gorzowski* od *Gorzów* a nie od *Gorzów Wielkopolski*).

Czasami niemożliwe okazywało się wskazanie związku natury morfologicznej pomiędzy przymiotnikiem relacyjnym a nazwą, do której się on odnosi. Ustalenie bazy derywacyjnej umotywowanej semantycznie oraz jej typu stawało się wówczas sporym problemem. W wypadku przymiotnika *angielski*, z uwagi na potencjalną metonimię (por. p. 9.4.1), nie zawsze wiadomo było, czy odnosi się on do oficjalnej nazwy kraju, tj. *Wielkiej Brytanii* czy też regionu – *Anglii*. Podobny problem dotyczył również przymiotnika *amerykański*. Czasami kontekst nie pozwalał odróżnić użyć odnoszących się do *Stanów Zjednoczonych Ameryki Północnej*, które potocznie nazywa się *Ameryką* (9.68), od użyć, które mogą wskazywać także na kontynent (9.69).

(9.68) Wydał także 2 zbiory opowiadań i około 30 opowiadań w angielskich, [amerykańskich]$^{\text{amerykański, }\textbf{Stany Zjednoczone}}_{\text{relAdj(}\textbf{placeName.country}\text{)}}$ i kanadyjskich czasopismach.

(9.69) Magazyn podał w środę wieczorem czasu [amerykańskiego]$^{\text{amerykański, }\textbf{Ameryka}}_{\text{relAdj(}\textbf{geogName}\text{)}}$ [...].

Jednak ustalenie obiektu odniesienia nie kończyło problemów związanych z anotacją przymiotników relacyjnych. Także podanie formy podstawowej niekiedy wymagało arbitralnych rozstrzygnięć. W wypadku państw formami podstawowymi mogą być ich oficjalne nazwy, np. *Zjednoczone Królestwo Wielkiej Brytanii i Irlandii Północnej*, *Republika Czeska*, *Stany Zjednoczone Ameryki Północnej*, albo też nazwy używane zwyczajowo, np. *Wielka Brytania*, *Czechy*, *Stany Zjednoczone/USA*. Poniżej podajemy przykłady normalizacji niektórych baz derywacyjnych zastosowanych w projekcie (pod warunkiem wcześniejszego ustalenia ich typu jako `placeName.country`). Są to najczęściej krótsze, zwyczajowe warianty nazw:

(9.70) *amerykański* ← *Stany Zjednoczone*

(9.71) *angielski*, *brytyjski* ← *Wielka Brytania*

(9.72) *południowoafrykański* ← *Republika Południowej Afryki*

(9.73) *sowiecki*, *radziecki* ← *Związek Radziecki*

Wreszcie postać bazy derywacyjnej mogła być kłopotliwa do ustalenia w przypadku, gdy pojawiały się nazwy w języku obcym. Według przyjętego przez nas założenia bazę derywacyjną oraz formę podstawową podawano wówczas również w języku obcym, jak w przykładach (9.74)–(9.75).

(9.74) [Sinfonietta [Cracovia]$^{\textbf{Cracovia}}_{\text{placeName.settlement}}$]$^{\text{Sinfonietta Cracovia}}_{\text{orgName}}$

(9.75) [[British]$^{\text{british, }\textbf{Great Britain}}_{\text{relAdj(placeName.country)}}$ Council]$^{\text{British Council}}_{\text{orgName}}$

Kolejne zagadnienie związane z anotacją przymiotników relacyjnych dotyczy ustalenia typu obiektu. Jak to opisywano już w p. 9.4.3, dla przymiotnika *europejski* możliwe są dwie różne bazy derywacyjne: *Europa* lub *Unia Europejska*, oraz trzy ich typy: nazwa geograficzna, blok lub organizacja.

Warto również wspomnieć o szczególnym przypadku przymiotników relacyjnych, które mogą zawierać się w swojej własnej bazie derywacyjnej, np. przymiotnik *świętokrzyski* może mieć za podstawę nazwę góry *Święty Krzyż* – zob. (9.76). Jednak częstszą i w wielu przypadkach bardziej uzasadnioną interpretacją będzie uznanie za bazę derywacyjną *województwa świętokrzyskiego* – zob. (9.77). Otwartą kwestią pozostaje jednak wówczas, jaka baza derywacyjna jest najbardziej uzasadniona w nazwie samego województwa – zob. (9.78).

(9.76) Chodzi o inwestycje w dziesięciu gminach z rejonu Gór [Świętokrzyskich]$^{\text{świętokrzyskie, }\textbf{Święty Krzyż}}_{\text{relAdj(geogName)}}$.

(9.77) pracownik [Świętokrzyskiego]$^{\text{świętokrzyski, }\textbf{województwo świętokrzyskie}}_{\text{relAdj(placeName.region)}}$ Ośrodka Doradztwa Rolniczego w Modliszewicach

(9.78) 86 milionów złotych [...] chce wytargować z unijnego Funduszu Spójności [...] województwo [świętokrzyskie]$^{\text{świętokrzyskie, }\textbf{Święty Krzyż}?}_{\text{relAdj(geogName)}}$.

9.4.5. Grupy religijne i ich członkowie

Nazwy oznaczające grupy i organizacje religijne oraz zakony zaliczone zostały w projekcie NKJP (jak wszystkie skupiska ludzkie) do typu `orgName`, natomiast nazwy ich członków były oznaczane jako ich derywaty, czyli nazwy typu `orgName` z atrybutem `derivType` przyjmującym wartość `persDeriv`. Jednak niekiedy ustalenie, o który z tych przypadków chodzi, sprawiało spore trudności. Zdarza się bowiem, że nazwa członków organizacji jest pospolicie używana na oznaczenie całej grupy, np. *karmelici*, *paulini*, *franciszkanie*, *benedyktyni*, *szyici*, *świadkowie Jehowy*. Nazwy takie mogły być zatem interpretowane dwojako. Niekiedy kontekst pozwalał rozstrzygnąć sporne kwestie, np.:

(9.79) Jako przedstawiciele [Świadków Jehowy]$^{\text{Świadkowie Jehowy}}_{\textbf{orgName}}$, są oczywiście całkowicie przeciwni przedmałżeńskim stosunkom.

(9.80) My jesteśmy [świadkami Jehowy]$^{\text{świadkowie Jehowy; Świadkowie Jehowy}}_{\textbf{persDeriv(orgName)}}$, a on wyznawał satanizm.

W przypadkach takich jak (9.80), o ile było to możliwe, podawana była także wartość atrybutu `derivedFrom`, np.:

(9.81) W X w. założył on w Trzemesznie klasztor [benedyktynów]$^{\text{benedyktyni; }\textbf{Zakon Świętego Benedykta}}_{\text{persDeriv(orgName)}}$.

W obrębie nazw organizacji religijnych i ich członków ciekawym przypadkiem jest *Krzyżak*. Oznacza on nie tylko członka zakonu, ale i obywatela państwa krzyżackiego. Ta nadwyżka znaczeniowa została zresztą utrwalona przez ortograficzny zapis wielką literą. Teoretycznie więc obok wspomnianych wyżej interpretacji, odzwierciedlonych w przykładach (9.82), (9.83) i (9.84), mogła pojawić się jeszcze jedna – jak w przykładzie (9.85):

(9.82) Na wojnę z [Krzyżakami]$^{\text{Krzyżacy}}_{\textbf{orgName}}$ wysłał Kórnik 2 żołnierzy.

(9.83) Odetchnęły zbuntowane przeciwko [Krzyżakom]$^{\text{Krzyżacy; }\textbf{Zakon Krzyżacki}}_{\textbf{persDeriv(orgName)}}$ dwory.

(9.84) [Krzyżacy]$^{\text{Krzyżacy; }\textbf{Zakon Szpitala Najświętszej Marii Panny Domu Niemieckiego w Jerozolimie}}_{\textbf{persDeriv(orgName)}}$ zmasakrowali miejscową ludność.

(9.85) [Krzyżacy]$^{\text{Krzyżak; }\textbf{Państwo Zakonu Krzyżackiego}}_{\textbf{persDeriv(placeName)}}$

Wydaje się jednak mało prawdopodobne, by nazwy Krzyżacy użyto w odniesieniu do obywateli państwa zakonnego, którzy nie są zakonnikami, dlatego ostatecznie w projekcie ograniczono się do anotacji (9.82)–(9.84). Jak pokazują przykłady (9.83) i (9.84), dla uzyskania spójności oznaczeń konieczne będzie w przyszłości ustalenie kanonicznych baz derywacyjnych dla tego typu jednostek, jak to zostało zrealizowane w wypadku nazw państw (zob. p. 9.4.4).

W wypadku niektórych organizacji i związków wyznaniowych trudno wyznaczyć granicę między nazwami własnymi a pospolitymi. Na granicy znajdują się jednostki typu:

(9.86) buddyzm, konfucjanizm, mahometanizm

Mimo że związek z nazwą własną, będącą dla tego typu jednostek bazą derywacyjną, jest silny (*buddyzm* pochodzi od *Budda*, a *mahometanizm* od *Mahomet*), a sama nazwa de facto oznacza także (metonimicznie) grupę religijną, wyrazy tego typu upowszechniły się w języku na tyle mocno, że odczuwane są na ogół jako pospolite, a nie nazwy własne, dlatego nie były anotowane.

9.4.6. Niepewny zasięg tekstowy nazw

Ustalanie lewych i prawych granic wystąpień tekstowych jednostek nazewniczych jest znanym problemem w automatycznej identyfikacji tych jednostek. Również w wypadku anotacji ręcznej, jak to ma miejsce w projekcie NKJP, zasięg danej jednostki w tekście nie zawsze jest łatwy do wskazania. W wypadku dat składniki słowne, jak *dzień, godzina, rok, stulecie, era* itd., uznawaliśmy za integralne części jednostek nazewniczych – por. (9.87). Tak nie było jednak w wypadku poprzedzających je przyimków – por. (9.88), mimo że oba wyrażenia jako okoliczniki są w wielu kontekstach wymienne.

(9.87) [**dnia** 15 maja 2002 **r.**]$^{\text{2002-05-15}}_{\text{date}}$

(9.88) **w** [dniu 15 maja 2002 r.]$^{\text{2002-05-15}}_{\text{date}}$

Dla nazw organizacji granice anotowanych jednostek powinny pokrywać się z oficjalnymi, urzędowymi nazwami. Dotyczy to w szczególności nazw miejscowych, które mogą, ale nie muszą, być częścią nazw obiektów geograficznych, instytucji, urzędów czy organizacji. W przykładzie (9.89) nazwa miasta jest niewątpliwie częścią nazwy budynku. Jednak ustalenie nazwy oficjalnej nie zawsze było proste, jak to ilustrują wahania w przykładach (9.90)–(9.91).

(9.89) w [Muzeum Narodowym
w [Warszawie]$^{\text{Warszawa}}_{\text{placeName.settlement}}$]$^{\text{Muzeum Narodowe w Warszawie}}_{\text{geogName}}$

(9.90) w [Centrum Handlowym]$^{\text{Centrum Handlowe}}_{\text{geogName}}$
w [Czeladzi]$^{\text{Czeladź}}_{\text{placeName.settlement}}$

(9.91) w [Centrum Handlowym
w [Czeladzi]$^{\text{Czeladź}}_{\text{placeName.settlement}}$]$^{\text{Centrum Handlowe w Czeladzi}}_{\text{geogName}}$

Podobny problem dotyczy przymiotników utworzonych od nazw miejscowości. Częstokroć przymiotnik taki stanowi część nazwy instytucji, firmy lub przedsiębiorstwa – por. (9.92), niekiedy jednak nie wchodzi w jej skład – por. (9.93).

(9.92) prezes [[Warszawskiego]$^{\text{warszawski, Warszawa}}_{\text{relAdj(placeName.settlement)}}$
Klubu Narciarskiego]$^{\text{Warszawski Klub Narciarski}}_{\text{orgName}}$

(9.93) Jednocześnie [warszawskie]$^{\text{warszawski, Warszawa}}_{\text{relAdj(placeName.settlement)}}$
[Muzeum Narodowe]$^{\text{Muzeum Narodowe}}_{\text{orgName}}$ pokazuje wystawę prac Christo i Jeanne-Claude.

W ustaleniu granic nazw oczywiście pomocna okazywała się ortografia. Przymiotniki pisane wielką literą na ogół są częścią nazwy, pisane małą leżą poza jej granicami. Trzeba jednak pamiętać, że NKJP jest zróżnicowanym zbiorem, zawierającym także teksty, w których nie przestrzega się w sposób rygorystyczny zasad ortograficznych. Niekiedy więc ustalenie granic było niemożliwe.

Zauważmy również, że w większości wypadków wyrazy pisane małą literą i wskazujące typ nazwanego obiektu, takie jak *firma*, *cmentarz* itd., nie należą do nazwy, nie powinny być zatem anotowane – zob. (9.94). Od tej reguły są jednak istotne i systematyczne wyjątki, np. w wypadku województw i gmin – zob. (9.101)–(9.102).

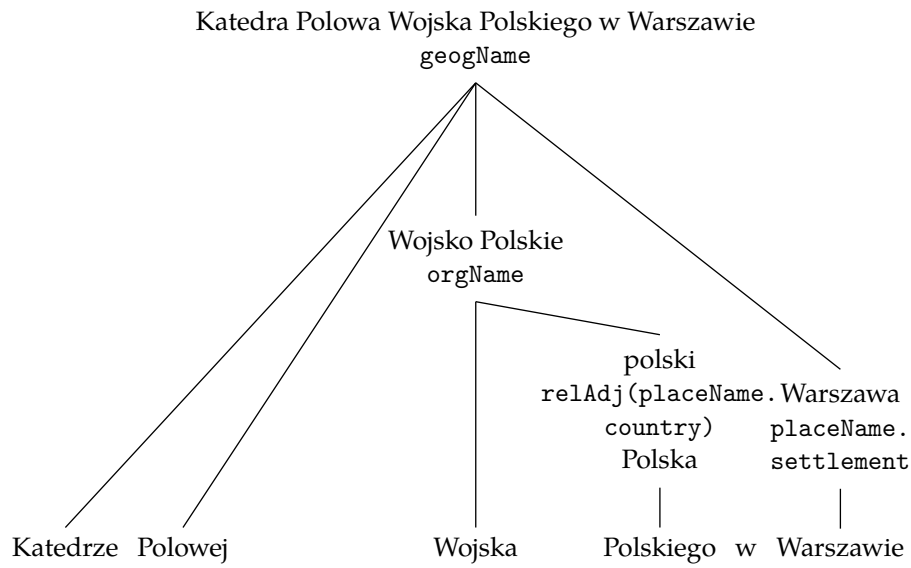
(9.94) cmentarz parafialny w [Dzierzgowie]$^{\text{Dzierzgów}}_{\text{placeName.settlement}}$

Procedura ustalenia oficjalnej nazwy niekiedy okazywała się zbyt czasochłonna, dlatego niekiedy granice nazwy wyznaczano intuicyjnie. Jednostki, w których wypadku pojawiły się problemy z ustaleniem zasięgu tekstowego, miały być opatrywane odpowiednim komentarzem.

9.4.7. Niepewny stopień zagnieżdżenia nazw

Jak już wspomniano, w projekcie NKJP zostały opisane całe nazwy własne oraz części składowe tych nazw (zagnieżdżenia), które spełniają omówione wcześniej kryteria jednostek anotowanych. Istnieją zatem często drzewa anotacyjne kilkupoziomowe i zawierające różne rodzaje anotowanych jednostek – zob. (9.95). Jest tak w szczególności dla jednostek zawierających nazwy osób – zob. (9.55), które to nazwy zgodnie z założeniami najczęściej otrzymują drzewa co najmniej trzypoziomowe.

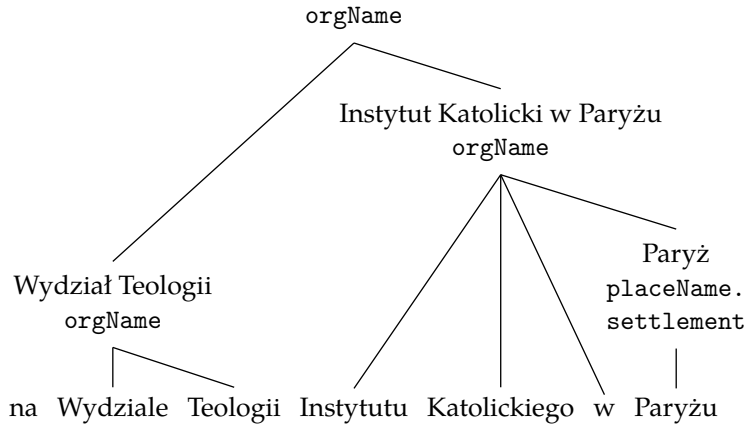
(9.95)



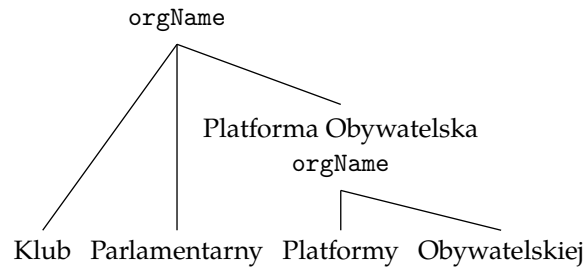
Niejaką trudność sprawiała anotacja złożonych nazw organizacji i decyzja o wyodrębnieniu ich części jako zagnieżdżonych. Jeśli element nazwy stanowił sam w sobie jednostkę instytucjonalną, będącą wydzieloną częścią nadrzędnej organizacji, anotowaliśmy ją jako taką (9.96), w przeciwnym wypadku pozo-

stawialiśmy nieoznaczoną. W niektórych przykładach to rozgraniczenie mogło być trudne, np. w (9.97), mimo podobieństwa struktury składniowej do przypadku (9.96), klub parlamentarny nie jest osobną organizacją w obrębie partii, a jedynie wskazaniem na tych członków parlamentu, którzy należą do klubu powołanego przez daną partię. Dlatego nie oznaczaliśmy tych dwóch członów jako zagnieżdżenia.

(9.96) Wydział Teologii Instytutu Katolickiego w Paryżu



(9.97) Klub Parlamentarny Platformy Obywatelskiej

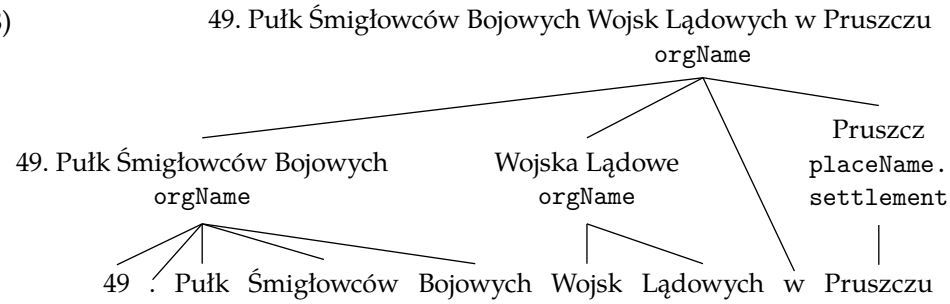


Czasem mylące może być zestawienie zagnieżdżenia z elipsą. W przykładach (9.98)–(9.100) ten sam ciąg *49. Pułk Śmigłowców Bojowych* był anotowany na trzy różne sposoby. W (9.98) jest on częścią nazwy organizacji nadrzędnej, w (9.99) jest osobną nazwą organizacji, a w (9.100) nie jest anotowany ani osobno, ani jako zagnieżdżenie. Przyczyną tych różnic jest właśnie elipsa. Tylko przypadek (9.98) opisuje pełną nazwę tej jednostki wojskowej[11]. W użyciu (9.99) jest ona formą eliptyczną tej pełnej formy (z pominięciem organizacji nadrzędnej), a w (9.100) jest elipsą, zachowującą jednakże nazwę lokalizacji. Ten ostatni przykład jest zatem

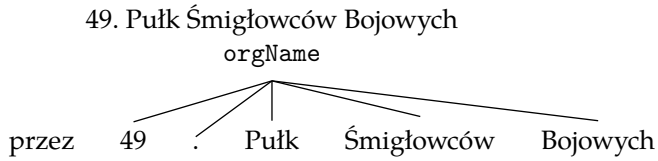
[11] W istocie oficjalna nazwa tej jednostki jest jeszcze bardziej złożona: *49. Pułk Śmigłowców Bojowych Wojsk Lądowych Rzeczpospolitej Polskiej w Pruszczu Gdańskim.*

podobny do (9.97), gdyż pułk ten nie jest częścią Pruszcza, w sensie organizacyjnym, podobnie jak klub parlamentarny nie jest częścią partii.

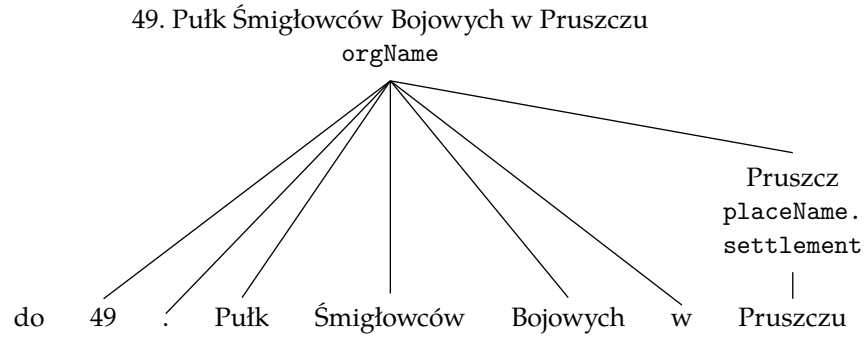
(9.98)



(9.99)



(9.100)



Kłopotliwe dla opisu zagnieżdżeń były też formy metonimiczne. W takich przykładach jak (9.101)–(9.102) *Stupsk* jest pierwotnie nazwą miejscowości, dlatego skłanialiśmy się do oznaczania go jako nazwy zagnieżdżonej (9.101), podobnie jak robiliśmy to z nazwą *Porto* w przykładzie (9.53) oraz *Paryż* w (9.96). Jednakże mogliśmy również uznać, że chodzi tak naprawdę o użycie metonimiczne, w którym słowo *Stupsk* byłoby nie zagnieżdżeniem, ale elipsą pełnej formy *gmina Stupsk* (9.102). Z uwagi na te wątpliwości w korpusie występują obecnie oba sposoby opisu.

(9.101) w [gminie [Stupsk]$^{\text{Stupsk}}_{\text{placeName.settlement}}$]$^{\text{gmina Stupsk}}_{\text{placeName.region}}$

(9.102) w [gminie Stupsk]$^{\text{gmina Stupsk}}_{\text{placeName.region}}$

Reasumując, do technik stosowanych przy wydzielaniu zagnieżdżeń należał test na elipsę i metonimię. Jeśli dany podciąg jest elipsą całej nazwy, a nie odnosi

się do instytucjonalnie wydzielonej części jakiejś organizacji, nie anotowaliśmy go jako zagnieżdżenia – por. (9.100). Jeśli dana nazwa jest metonimią innej, to tej drugiej również nie oznaczaliśmy jako zagnieżdżenia, dlatego *Europa* jako nazwa geograficzna nie pojawia się w anotacji *Europy* jako Unii Europejskiej – zob. (9.42). Należało jednak uważnie analizować przypadki zawierające m.in. opisy miejsc, w których nazwa może być wynikiem kombinacji zagnieżdżenia, elipsy i metonimii, takie jak (9.50)–(9.51), (9.53), (9.55)–(9.56), (9.101)–(9.102).

Zauważmy na koniec, iż w projekcie NKJP nie były opisywane zagnieżdżenia dla baz derywacyjnych. Dlatego np. w nazwie z przykładów (9.83)–(9.84) nie precyzowano, od czego pochodzą części składowe baz derywacyjnych *Krzyżacki* i *Niemieckiego*. Dzięki temu unikamy m.in. „zapętlania" się opisu w przypadkach takich jak (9.43)–(9.45) czy (9.77).

9.4.8. Zmienna szczegółowość skali dla wyrażeń czasu

Wyrażenia służące określaniu czasu pozwalają opisywać bardzo różnorodne zjawiska. Umożliwiają nie tylko wskazanie momentu jakiegoś zdarzenia w sposób precyzyjny, przez umieszczenie go na bezwzględnej osi czasu, jak w przykładzie (9.87), ale też w sposób niedookreślony, względny w stosunku do momentu wypowiedzi (np. *w zeszłym roku*) lub do innej daty czy wydarzenia (np. *miesiąc po Nowym Roku*). Jak wspomniano wcześniej, w projekcie NKJP anotacja jednostek wyrażających czas została uproszczona i ograniczona do dat określonych w sposób bezwzględny, które można było opisać za pomocą jednego z dwóch typów: `date` oraz `time`.

Oznaczane wyrażenia odnosiły się do jednostek o różnej długości trwania. Jeżeli tylko było to możliwe, tj. w wypadku jednostek wskazujących na konkretny moment w skali od sekundy do roku, określany był również atrybut `when`:

(9.103) Środa jest dla Brighwella ważnym dniem; o [godzinie jedenastej, 11 minut i 11 sekund]$^{11:11:11}_{\text{time}}$, po raz jedenasty w życiu zobaczy całkowite zaćmienie Słońca.

(9.104) Mściwój II (zmarł w [1294 r.]$^{1294}_{\text{date}}$) nie miał syna.

Zgodnie z przyjętymi w NKJP standardami, w wypadku wyrażeń odnoszących się do dłuższych okresów, np. dekad, stuleci, anotacja ograniczała się do wskazania konkretnych dat. Pole zawierające atrybut `when` pozostawało puste, np.:

(9.105) w końcu [lat osiemdziesiątych]$_{\text{date}}$

(9.106) około [XV wieku p.n.e]$_{\text{date}}$

(9.107) w drugiej połowie [XVI w.]$_{\text{date}}$

9.4.9. Problematyka geopolityczna i narodowościowa

W trakcie prac nad oznaczaniem nazw własnych pojawił się szereg pytań dotyczących klasyfikowania nazw geograficznych, a także powiązanych z nimi nazw narodowości. Często odpowiedź na takie pytania wymagała ustalania dodatkowych kryteriów odwołujących się do wiedzy pozalingwistycznej. Wielokrotnie trzeba było też podjąć arbitralne decyzje co do zaszeregowania danej nazwy (lub jej bazy derywacyjnej) w ramach klasyfikacji nazw geograficznych i geopolitycznych przyjętej w NKJP.

W tekstach korpusu dość często pojawiają się nazwy państw. W ramach klasyfikacji odpowiada im kategoria `placename.country` obejmująca państwa, kraje, kolonie i wspólnoty, a także narody je zamieszkujące:

(9.108) [Polska]$^{\text{Polska}}_{\textbf{placeName.country}}$ dla [Polaków]$^{\text{Polak; Polska}}_{\textbf{persDeriv(placeName.country)}}$

W trakcie oznaczania napotkano szereg nazw budzących wątpliwość, czy należy je uznać za państwa, np.:

(9.109) Palestyna, Czeczenia, Abchazja, Kosowo

Jako kryterium rozstrzygające przyjęto, czy dane państwo jest oficjalnie uznane przez Polskę oraz przez instytucje międzynarodowe – stąd decyzja o zaklasyfikowaniu wyżej wymienionych nazw jako `placename.region`. Jednakże mapa polityczna świata zmienia się dynamicznie. W trakcie projektu Międzynarodowy Trybunał Sprawiedliwości w Hadze przyjął 22 lipca 2010 rezolucję uznającą deklarację niepodległości Kosowa. Zgodnie z przyjętymi założeniami, w tekstach dotyczących okresów przed tą datą *Kosowo* oznaczane jest jako region – zob. przykład (9.110), a w tekstach na temat późniejszych okresów jako kraj – zob. (9.111).

(9.110) Rozważa się również możliwości współpracy w budowie elektrociepłowni w [kosowskim]$^{\text{kosowski; Kosowo}}_{\text{relAdj(placeName.}\textbf{region}\text{)}}$ zagłębiu węglowym w Jugosławii.

(9.111) Decyzję o likwidacji misji podjęli po wizytacji baz w Albanii i [Kosowie]$^{\text{Kosowo}}_{\text{placeName.}\textbf{country}}$ przedstawiciele Komendy Głównej Państwowej Straży Pożarnej i MSWiA.

Możliwa jest również sytuacja odwrotna – np. Brandenburgia stanowiła kiedyś osobny organizm państwowy, natomiast obecnie jest częścią składową Republik Federalnej Niemiec. W zależności od kontekstu historycznego tekstu, o ile był możliwy do ustalenia, oznaczaliśmy ją odpowiednio jako kraj – zob. (9.112), region – zob. (9.113) lub też krainę geograficzną. Do tych intrepretacji dochodziła zwyczajowo również taka, w której nazwa jednostki podziału administracyjnego określa w istocie jej mieszkańców – zob. (9.114).

(9.112) 1618 traktat polsko-[brandenburski]$^{\text{brandenburski; Brandenburgia}}_{\text{relAdj(placeName.}\mathbf{country}\text{)}}$, kładący kres wojnie celnej na szlaku warciańsko-odrzańskim [...].

(9.113) Również cztery lata później odrzucił propozycję robienia kariery w Berlinie i objął funkcję premiera [Brandenburgii]$^{\text{Brandenburgia}}_{\text{placeName.}\mathbf{region}}$.

(9.114) Z tego powodu cierpiało całe Pomorze, obawiając się – po śmierci księcia – najazdu [Brandenburgii]$^{\text{Brandenburgia}}_{\mathbf{orgName}}$ lub zakonu krzyżackiego.

Rzadziej oznaczaną kategorią była `placename.bloc`, do której zaliczamy nazwy organizmów złożonych z dwu lub więcej państw. Pojawia się tu problem klasyfikacji państw federacyjnych. Możliwe jest zaklasyfikowanie federacji jako bloku, a jej składowych jako krajów, lub też federacji jako kraju, a składowych jako regionów. Za czynniki decydujące przy wyborze przyjęto, oprócz samego formalnego ustroju (konstytucji), ocenę rzeczywistej autonomii poszczególnych regionów i uznania ich części składowych za państwa, a także osąd, czy silniejsza jest identyfikacja ogółu obywateli z mniejszą czy większą jednostką jako państwem. Biorąc pod uwagę te czynniki, *Związek Radziecki* był oznaczany jako kraj, jak w (9.115)–(9.116), a jego republiki składowe jako regiony, mimo że formalnie każda republika miała konstytucyjnie zagwarantowaną autonomię oraz prawo do opuszczenia federacji (por. artykuł 72 Konstytucji ZSRR 1978). Chociaż Ukraińska SRR i Białoruska SRR miały oddzielne reprezentacje w ONZ, to oznaczaliśmy je jako regiony. Z kolei *Unia Europejska* była oznaczana jako blok (zob. (9.117)), a jej państwa składowe jako kraje, z uwagi na silniejszą identyfikację obywateli z poszczególnymi państwami członkowskimi niż ze wspólnotą nadrzędną.

(9.115) Czołgi [radzieckie]$^{\text{radziecki; Związek Radziecki}}_{\text{relAdj(placeName.}\mathbf{country}\text{)}}$ wjechały do Rygi 17 czerwca 1940 r. w samo południe.

(9.116) Trudno mówić o przyjemności podróżowania [sowieckimi]$^{\text{sowiecki; Związek Radziecki}}_{\text{relAdj(placeName.}\mathbf{country}\text{)}}$ drogami.

(9.117) W jednorodnej [Unii [Europejskiej]$^{\text{europejski; Europa}}_{\text{placeName.bloc}}$]$^{\text{Unia Europejska}}_{\text{placeName.}\mathbf{bloc}}$ jest wiele terytoriów o specjalnym statusie, takich jak Gibraltar, Madera, góra Athos, Wyspy Owcze, Laponia, Azory czy Wyspy Alandzkie.

(9.118) Do [Unii [Europejskiej]$^{\text{europejski; Europa}}_{\text{placeName.bloc}}$]$^{\text{Unia Europejska}}_{\text{placeName.}\mathbf{bloc}}$ przyjęto de facto część [grecką]$^{\text{grecki; Grecja}}_{\text{relAdj(}\mathbf{placeName.country}\text{)}}$ (Cypru).

Przy oznaczaniu narodowości przyjęto ogólną zasadę relacji między nazwą państwa a określeniem obywatela dane państwo zamieszkującego, jak to pokazano na przykładzie (9.108). Jest to związek naturalny dla większości państw europejskich, które były tworzone jako państwa narodowe i stąd zakorzeniony w tradycji języka polskiego. Stanowi natomiast daleko idące uproszczenie

skomplikowanej mozaiki zależności między nazwami państw, narodów, języków oraz wyznawanych religii. Szczegółową analizę współzależności tych elementów podaje Szul (2009). Dla wielu państw na świecie związek między narodem a państwem nie jest jednoznaczny. Istnieją państwa wielonarodowe, gdzie głównym czynnikiem identyfikacyjnym jest religia. Na przykład – zob. (9.119) – obywatel Indii to Hindus, czyli najczęściej wyznawca religii hinduistycznej (hindus), choć kraj ten zamieszkuje w istocie wiele różnych narodowości.

(9.119) Krzemowe miasteczka tworzą lub zamierzają tworzyć Anglicy, Francuzi, $[\text{Hindusi}]^{\text{Hindus; } \textbf{Indie}}_{\text{persDeriv}(\textbf{placeName.country})}$, Izraelczycy.

Zdarza się sytuacja odwrotna, gdy dany naród wytworzył wiele państw, ale jest określany w języku polskim jedną nazwą. Przy oznaczaniu przyjęto w tej sytuacji dwie metody. I tak Albańczycy, którzy utworzyli dwa państwa narodowe, Albanię i Kosowo, byli w korpusie przypisywani do Albanii – zob. (9.120) – choć nie jest wykluczone, że w przyszłości ukonstytuuje się odrębny naród Kosowa i uzyska własne określenie w języku polskim. Drugi przypadek to Arabowie, którzy zamieszkują kilka państw Bliskiego Wschodu i Afryki Północnej. W korpusie przyjęto oznaczenie tej społeczności bez wskazania państwa, od którego ona pochodzi – zob. (9.121).

(9.120) Serbowie i $[\text{Albańczycy}]^{\text{Albańczyk; } \textbf{Albania}}_{\text{persDeriv(placeName.country)}}$ z Kosowa podpisali w poniedziałek porozumienie o szkolnictwie [...].

(9.121) W czasie tym Iran, wyczerpany, wyniszczony wiekowymi wojnami z Bizancjum, jest świeżo podbity przez $[\text{Arabów}]^{\text{Arab; \_}}_{\text{persDeriv(placeName.country)}}$, którzy zaczynają krzewić nową wiarę – islam.

Istnieją też narody, które nigdy nie wytworzyły własnego państwa i nie są związane z określonym terytorium, np. Romowie, których nazwy oznaczaliśmy jako derywacje osobowe bez określenia bazy derywacyjnej – por. (9.122). Podobne podejście przyjęto w stosunku do Żydów, którzy są narodem bardzo starym, w ciągu wieków żyjącym w diasporze i w swej historii związanym silniej z religią niż ideą państwowości. Naród ten wytworzył jednak własne państwo i określenie obywatela tego państwa, Izraelczyk, oznaczamy jako pochodzące od Izraela (zob. (9.123)).

(9.122) Chodzi przecież o sprawienie, żeby ateista, $[\text{Żyd}]^{\text{Żyd; \_}}_{\text{persDeriv(placeName.country)}}$, [...] $[\text{Cygan}]^{\text{Cygan; \_}}_{\text{persDeriv(placeName.country)}}$ czy $[\text{Arab}]^{\text{Arab; \_}}_{\text{persDeriv(placeName.country)}}$ zakładał łatwo przedsiębiorstwo.

(9.123) Myślę, że to są często $[\text{Izraelczycy}]^{\text{Izraelczyk; } \textbf{Izrael}}_{\text{persDeriv(placeName.country)}}$ i często $[\text{Żydzi}]^{\text{Żyd; \_}}_{\text{persDeriv(placeName.country)}}$, nie można jednoznacznie odpowiedzieć.

Jako derywacje osobowe, nieprzypisane do konkretnej bazy derywacyjnej, oznaczano też nazwy przedstawicieli szerszych grup etnicznych niezwiązanych bezpośrednio z narodem, np. *Indianin* oraz synonimy *Murzyn* i *Afroamerykanin* (skomplikowane związki etymologiczne odpowiednio z Indiami i Afryką/Ameryką nie dałyby się prosto odzwierciedlić w pojedynczej bazie derywacyjnej).

Jak już wspomniano, określenia metaforyczne pochodzące pośrednio od nazw narodowości i grup etnicznych, często pisane małą literą, nie były w ogóle oznaczane jako jednostki nazewnicze (np. *francuz* 'klucz francuski' czy *murzyn* 'osoba wykonująca pracę za kogoś'). Nie oznaczaliśmy też pejoratywnych określeń grup etnicznych i narodów pisanych małą literą (np. *bambo*, *żółtek*, *żabojad*).

9.4.10. Nazwy błędne i kwestie nierozstrzygnięte

Przypomnijmy, że poziom nazw własnych opiera się w NKJP na poziomie morfoskładniowym, który z kolei tworzony jest na podstawie poziomu segmentacji. Z powodu pewnych decyzji podjętych w pierwszej fazie projektu i dotyczących właśnie tego najniższego poziomu, założenia dotyczące nazw własnych, opisane w p. 9.3, nie zawsze mogły być precyzyjnie stosowane. Na przykład nazwiska zawierające dywiz (-) nie zawsze dzielone były na kilka segmentów – por. (9.124). Niemożliwe było wówczas opisanie każdego z nazwisk składowych, zanim połączone one zostały przez wspólny węzeł. Podobny problem dotyczył skrótów takich jak w (9.125), w których w idealnym przypadku każda z dwóch pierwszych liter powinna być opisana jako inicjał imienia (*Jan Paweł*), co jest jednak niemożliwe z uwagi na istnienie tu tylko jednego segmentu.

(9.124) komisarz $[[\text{Danuta}]^{\text{Danuta}}_{\text{persName.forename}}$
$\textbf{[Wołk-Karczewska]}^{\text{Wołk-Karczewska}}_{\text{persName.surname}}]^{\text{Danuta Wołk-Karczewska}}_{\text{persName}}$

(9.125) Przy całym szacunku dla $\textbf{[JPII]}^{\text{JPII}}_{\text{persName}}$ jako człowieka [...].

Inne typy błędów, np. ortograficzne, mogą pochodzić bezpośrednio z tekstu źródłowego. Jeśli występują one w jednostkach nazewniczych, staramy się podawać ich poprawną formę podstawową – zob. (9.126).

(9.126) Czego nie umiała dokonać socjal-demokratyczna lewica zachodu, tego dokonała partia bolszewików, partia Włodzimierza $\textbf{[Lenia]}^{\textbf{Lenin}}_{\text{persName.surname}}$.

Dla niektórych nazw nie udało się ustalić poprawnej anotacji ani na podstawie kontekstu, ani na podstawie badań pozakorpusowych (np. Internet, ogólna wiedza anotatora); do takich kwestii należą problemy z niektórymi nazwami miejscowości i wyrażeniami czasowymi.

Istnieją w języku polskim nazwy miejscowości różniące się formą mianownikową, ale identyczne w przypadkach zależnych (poza biernikiem), np. *Stryków/Strykowo*, *Szczodrów/Szczodrowo*, *Kunów/Kunowo*. W niektórych zdaniach

(zob. (9.127)) niemożliwe było zatem jednoznaczne wskazanie formy podstawowej, co anotator mógł zasygnalizować przez wybór niższego stopnia pewności anotacji (`@cert`) oraz podanie możliwych form alternatywnych w komentarzu (`@certComment`). Ten sam problem dotyczył również ustalania baz derywacyjnych dla przymiotników i nazw mieszkańców, np. *kunowianin*.

(9.127) Magdalena Kania z [Kunowa]$^{\textbf{Kunów}\text{ albo }\textbf{Kunowo}}_{\text{placeName.settlement}}$

Niemożność rozstrzygnięcia dotyczyła również wyrażeń czasowych, kiedy podana była godzina bez żadnych dodatkowych określeń typu: rano, wieczorem, po południu – por. (9.128).

(9.128) Pociąg pospieszny do Jakiegoś Tam odchodzi o [dziewiątej dwadzieścia]$^{\textbf{09:20:00}\text{ albo }\textbf{21:20:00}}_{\text{time}}$.

9.5. Wnioski i perspektywy

W niniejszym rozdziale zaprezentowaliśmy zasady dotyczące anotacji Narodowego Korpusu Języka Polskiego na poziomie jednostek nazewniczych. Zdefiniowaliśmy zakres anotacji obejmujący nazwy osób, organizacji, obiektów geograficznych i geopolitycznych, określenia dat i godzin, a także przymiotniki relacyjne, nazwy mieszkańców, narodów i członków odnoszące się do czterech pierwszych kategorii. Przedstawiliśmy następnie taksonomię tych jednostek (zawierającą sześć typów głównych i osiem podtypów) oraz zestaw przypisywanych im atrybutów. Podaliśmy również ważniejsze strategie anotacyjne, w tym m.in. dwa nowatorskie założenia: o anotacji wszystkich nazw zagnieżdżonych w danej jednostce nazewniczej oraz o rozłącznym oznaczaniu jednostek występujących w strukturach spójnikowych.

W drugiej części rozdziału dokonaliśmy analizy zjawisk napotkanych w trakcie anotacji korpusu, które utrudniają lub wręcz uniemożliwiają zastosowanie przyjętych wcześniej zasad anotacji. Należą do nich w pierwszej kolejności metonimia i elipsa, które mają wpływ na ustalanie typów i podtypów, a także na identyfikację stopnia zagnieżdżenia części składowych niektórych nazw. Szczególnie trudny okazuje się niekiedy wybór właściwych baz derywacyjnych dla przymiotników relacyjnych i narodów, co spowodowane jest m.in. aspektami pozajęzykowymi, takimi jak sporny status regionów i państw, brak pokrycia geograficznego między narodem a państwem itp.

Zaznaczmy, że wszystkie przykłady cytowane w niniejszym rozdziale pochodzą z korpusu NKJP.

Metodologia anotacji jednostek nazewniczych, według opisanych tu zasad, a także narzędzia używane w tym procesie, przedstawione są w rozdz. 13. Pierwsze wyniki ilościowe tej anotacji podane są w tab. 9.1. Nazwy osobowe

stanowią ponad połowę wszystkich jednostek, należy jednak pamiętać, że specyficzne reguły anotacyjne dotyczące ich stopnia zgnieżdżania niekiedy sztucznie zawyżają tę liczbę. Niewielki procent oznaczonych jednostek opatrzony

Tabela 9.1. Liczba wystąpień jednostek nazewniczych oznaczonych w NKJP (w wersji z 30 marca 2011)

Nazwy osobowe	**Nazwy organizacji**	**Nazwy geograficzne**	**Nazwy geopolityczne**
47286	11380	3893	10733
Daty	**Godziny**	**Przymiotniki relacyjne**	**Derywacje osobowe**
4514	562	7147	1785
Wszystkie jednostki nazewnicze			87300

został wskaźnikiem niepewności anotatora co do zaproponowanej przez niego anotacji. Mianowicie 133 jednostki (a więc 0,15% całości) mają w wersji końcowej (po superanotacji – zob. rozdz. 13) atrybut `@cert` o wartości `medium`, a 165 jednostek (0,19%) – o wartości `low`.

Dokonany dotychczas opis jednostek nazewniczych w NKJP powinien w przyszłości być uzupełniany i udoskonalany. Ważne byłoby wzięcie pod uwagę nowych kategorii, takich jak tytuły osób, nazwy produktów i wydarzeń, jednostek miar i wag, a także dogłębniejszego opisu wyrażeń czasowych (okresów, dat względnych i przybliżonych itp.). Mimo ścisłej procedury kontroli jakości, wspólnej dla wszystkich warstw korpusu (dwaj anotatorzy i jeden superanotator przeglądali każdy jego fragment), jakość anotacji zapewne zmieniała się wraz ze stopniem zaawansowania projektu. Dlatego celowe byłoby ponowienie weryfikacji zgodności wyników z założeniami anotacji, a także uzupełnienie niektórych z tych założeń. Na przykład wybór kanonicznych baz derywacyjnych wspomnianych w p. 9.4.4 moglibyśmy skonfrontować z publikacjami Komisji Standardyzacji Nazw Geograficznych[12]. Użyteczne mogłoby się okazać zdefiniowanie baz kanonicznych również dla nazw osób i organizacji. Dodatkowo wydaje się, że konieczne jest pewne uszczegółowienie zasad wydzielania zagnieżdżeń, co ilustrują wahania w przypadkach takich jak (9.50)–(9.51) i (9.101)–(9.102). Przydatne byłoby zarazem dopuszczenie wprowadzania anotacji alternatywnych, wówczas gdy nie daje się ustalić jedynej możliwej interpretacji, np. (9.67), (9.127) czy (9.128).

Opisana tu warstwa anotacji jednostek nazewniczych w NKJP może jednak już teraz stać się podstawą do ciekawych badań korpusowych. Chcielibyśmy dokonać analizy kontekstów, które warunkują użycia metonimiczne nazw, co mogłoby pomóc np. w poprawie systemów do ich automatycznego rozpoznawania

[12] `http://ksng.gugik.gov.pl/`.

i kategoryzacji, takich jak *SProUT* (por. rozdz. 13). Ciekawi nas również, w jakim stopniu segmenty zidentyfikowane w warstwie jednostek nazewniczych odpowiadają segmentom opisanym na poziomie grup składniowych (por. rozdz. 8). Ta odpowiedniość warunkuje bowiem m.in. analizę morfoskładniową jednostek nazewniczych (chcielibyśmy móc wskazać centrum składniowe i semantyczne danej nazwy, jej przypadek, liczbę i rodzaj itp.). Wiąże się z tym również rozpoczynający się właśnie projekt, który ma na celu dodanie nowej warstwy anotacyjnej do NKJP, a mianowicie anotacji nawiązań (czyli koreferencji). Nawiązania powinny prawdopodobnie odwoływać się do grup składniowych i ich centrów, podczas gdy analiza semantyczna tak opisanego dyskursu jest zapewne łatwiejsza przy użyciu warstwy jednostek nazewniczych. Jeszcze inną perpektywę otwiera znajomość form podstawowych jednostek nazewniczych. Ich powiązanie z warstwą morfoskładniową i składniową ma szansę pomóc w rozwoju metod automatycznej lematyzacji nazw wielowyrazowych (por. pracę Piskorskiego i in. 2007).

10

Znakowanie XML

Adam Przepiórkowski

10.1. Standardy znakowania XML korpusów językowych

W stworzonym w pierwszej połowie lat 2000. Korpusie IPI PAN (Przepiórkowski 2004) do reprezentacji tekstów wykorzystany został standard XCES (XML Corpus Encoding Standard; Ide i in. 2000). Standard ten wyewoluował z wcześniejszego SGML-owego standardu CES (Corpus Encoding Standard; Ide i Véronis 1993). XCES definiuje reprezentację metadanych, struktury tekstu, a także znakowania morfosyntaktycznego, lecz nie znakowania składniowego, dlatego też w wypadku NKJP okazało się konieczne znalezienie rozwiązania ogólniejszego.

W wyniku porównania proponowanych standardów reprezentacji tekstów (Przepiórkowski i Bański 2009), biorącego pod uwagę m.in. ich rozpowszechnienie, stabilność i możliwość reprezentacji wszystkich poziomów lingwistycznych zdefiniowanych w NKJP, wybrany został – jako w zasadzie bezkonkurencyjny – standard Text Encoding Initiative (TEI), a konkretnie aktualna piąta wersja standardu (TEI P5; Burnard i Bauman 2008). Jedną z zalet tego wyboru jest fakt, że semantyka wykorzystywanych elementów i atrybutów została dobrze zdefiniowana w dokumentacji TEI[1], dzięki czemu nie było konieczne tworzenie osobnej rozbudowanej dokumentacji schematu XML dla NKJP.

Głównym problemem z wykorzystaniem standardu TEI w NKJP okazało się jego bogactwo[2]. TEI jest uniwersalnym standardem reprezentacji różnorodnych dokumentów tekstowych, od manuskryptów historycznych i słowników po nagrania mowy i korpusy językowe, pozwalającym na zapisanie informacji o wszelkich cechach tekstów i ich fragmentów, konieczne więc było wybranie tych elementów TEI, które faktycznie są przydatne w NKJP. Wybór ten nie zawsze był

[1] Zob. `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html`.

[2] Wystarczy wspomnieć, że dokumentacja TEI P5 liczy ok. 1500 stron.

prosty, gdyż niektóre rodzaje informacji, choćby o konstrukcjach składniowych, mogą być w TEI reprezentowane na różne sposoby. W takich sytuacjach staraliśmy się wybrać te schematy reprezentacji, które są najbardziej podobne do innych standardów będących w użyciu (braliśmy pod uwagę m.in. rodzinę powstających standardów ISO, a także TIGER-XML, PAULA itp.). W efekcie wykonana została praca analogiczna do stworzenia wspomnianego powyżej standardu CES, powstałego na podstawie ówczesnej wersji TEI.

W serii artykułów Bańskiego i Przepiórkowskiego (Przepiórkowski i Bański 2009, 2011, Bański i Przepiórkowski 2009, 2010 oraz Przepiórkowski 2009b) zostały opisane i uzasadnione poszczególne rozwiązania[3]. Niektóre z nich nieznacznie ewoluowały od czasu ukazania się wspomnianych artykułów[4], jednak podana tam motywacja dla wybrania takich a nie innych rozwiązań jest aktualna. W niniejszym rozdziale skupimy się więc na opisie reprezentacji poszczególnych poziomów znakowania tekstu w końcowym stadium projektu NKJP, nie wdając się już w dyskusje na temat możliwych alternatyw.

W punktach 10.2–10.8 przedstawione zostały zasady znakowania poszczególnych poziomów stosowane w pełnej wersji NKJP. Drobne różnice w znakowaniu ręcznie anotowanego podkorpusu wielkości 1 miliona słów (dalej zwanego korpusem *milionowym* lub *1M*; zob. rozdz. 5) w stosunku do schematu przedstawionego w tych punktach zostały omówione w p. 10.9.

10.2. Reprezentacja tekstu w NKJP

Na reprezentację każdego tekstu w korpusie składa się kilka plików:

1. `header.xml`: nagłówek, zawiera m.in. informacje o pochodzeniu tekstu;
2. `text_structure.xml`: właściwy tekst, wraz z informacją o jego strukturze;
3. `ann_segmentation.xml`: reprezentacja segmentacji tekstu na zdania i na segmenty (w przybliżeniu: słowa ortograficzne);
4. `ann_senses.xml`: informacja o znaczeniach wybranych słów wieloznacznych;
5. `ann_morphosyntax.xml`: interpretacje morfosyntaktyczne poszczególnych segmentów;
6. `ann_words.xml`: słowa składniowe i ich interpretacje morfosyntaktyczne;
7. `ann_groups.xml`: grupy składniowe;
8. `ann_named.xml`: jednostki nazewnicze.

[3] Punkty 10.3–10.4 oparte są na artykule Przepiórkowskiego i Bańskiego (2011), a pozostałe w dużej części na artykule Przepiórkowskiego i Bańskiego (2009).

[4] Na przykład bogatsza niż opisana w artykule Przepiórkowskiego (2009b), ale jednocześnie bliższa poziomowi morfosyntaktycznemu, jest reprezentacja słów składniowych.

Struktura tych plików i typy informacji w nich reprezentowane zostaną omówione w kolejnych punktach.

Oprócz plików reprezentujących poszczególne teksty w korpusie znajdują się także dwa pliki globalne, właściwe dla całego korpusu, ale wykorzystywane w znakowaniu poszczególnych tekstów:

9. `NKJP_header.xml`: nagłówek korpusu (zob. p. 10.3.1);
10. `NKJP_WSI.xml`: słownik sensów słów (ang. *word sense inventory*; zob. p. 10.8.3).

10.3. Metadane

10.3.1. Nagłówek korpusu

Zgodnie ze standardem TEI, nagłówek NKJP składa się z czterech części zawartych w elemencie `<teiHeader xml:lang="en" type="corpus">`: `<fileDesc>`, `<profileDesc>`, `<encodingDesc>` oraz `<revisionDesc>`.

Dwie z tych części mają bardzo prostą strukturę. Zawartość elementu `<profileDesc>` wskazuje na główne języki używane w znakowaniu tekstu i metadanych; pełna specyfikacja tej części podana jest poniżej:

```
<profileDesc>
 <langUsage>
  <language ident="pl">Polish</language>
  <language ident="en">English</language>
 </langUsage>
</profileDesc>
```

Wartości atrybutów `@ident` mogą być użyte w dowolnym elemencie XML do zasygnalizowania języka, w którym podana jest zawartość danego elementu: po ich zdefiniowaniu w elemencie `<langUsage>` mogą one być wartościami atrybutów `@xml:lang`. Ściślej rzecz ujmując, specyfikacja `@xml:lang="en"` we wspomnianym powyżej elemencie `<teiHeader>` jest dziedziczona przez pozostałe elementy, przez co angielski jest językiem domyślnym w nagłówku korpusu, choć w poszczególnych elementach ta specyfikacja może zostać w sposób jawny zmieniona, poprzez dodanie atrybutu `@xml:lang="pl"`.

Drugą prostą i jednorodną częścią nagłówka korpusu jest `<revisionDesc>`: zawiera ona ciąg elementów `<change>` opisujących zmiany w nagłówku lub w strukturze korpusu, np.:

```
<change who="#adamp" when="2009-08-01">
  Added <gi>profileDesc</gi>.
</change>
```

Kolejna część nagłówka, <fileDesc>, składa się z czterech mniejszych części. Pierwsza z nich, <titleStmt>, zawiera informacje o nazwie korpusu i o instytucjach i osobach odpowiedzialnych za jego stworzenie, np.:

```
<respStmt>
 <persName xml:id="bansp">Piotr Bański</persName>
 <resp>initial design of various XML schemata</resp>
</respStmt>
```

Do identyfikatorów zdefiniowanych za pomocą @xml:id można się odnosić w innych częściach nagłówka; zob. np. who="#adamp" w powyższym przykładzie elementu <change>.

Pozostałe trzy części elementu <fileDesc> to: <editionStmt> – krótkie stwierdzenie o stopniu stabilności danej wersji, <publicationStmt> – dane dotyczące dystrybucji korpusu oraz <sourceDesc> – ogólne informacje o pochodzeniu tekstów w korpusie (szczegółowe informacje są zawarte w nagłówkach poszczególnych tekstów).

Największą częścią nagłówka jest <encodingDesc>[5], zawierająca różnorodną charakterystykę korpusu. Tak więc na przykład w <projectDesc> znajduje się opis projektu ze strony http://nkjp.pl/, w <samplingDecl> – informacja o próbkowaniu tekstów w korpusie, zaś w <editorialDecl> – o przyjętych zasadach anonimizacji tekstów mówionych i ew. innych interwencjach edytorskich w tekstach składających się na NKJP.

Podczas gdy elementy te zawierają swobodne opisy, wiele innych części elementu <encodingDesc> jest bardziej ustrukturyzowanych. Być może najważniejsza z nich to ciąg elementów <classDecl> zawierających klasyfikacje, do których odwołują się nagłówki poszczególnych tekstów. Na przykład jedną z klasyfikacji wykorzystywanych w NKJP jest Uniwersalna Klasyfikacja Dziesiętna (UKD), zdefiniowana w nagłówku korpusu w sposób następujący:

```
<classDecl>
 <taxonomy xml:id="ukd">
  <bibl>
   <title xml:lang="pl">Uniwersalna Klasyfikacja Dziesiętna</title>
   <title xml:lang="en">Universal Decimal Classification</title>
   <edition>UDC-P058</edition>
  </bibl>
 </taxonomy>
</classDecl>
```

[5] W niniejszym opisie tej części nagłówka pomijamy elementy <tagsDecl> i <refsDecl>; zob. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-encodingDesc.html.

W nagłówku konkretnego tekstu (zob. p. 10.3.2 poniżej), odnośnik do tej klasyfikacji może wyglądać następująco:

```
<classCode scheme="#ukd">821.162.1-3</classCode>
```

W wypadku UKD nagłówek nie zawiera całej taksonomii, gdyż jest ona przyjętym w bibliotekarstwie standardem, a jedynie definiuje wersję, która jest w projekcie wykorzystywana. Inaczej jest w wypadku stylistycznej klasyfikacji tekstów (por. rozdz. 2), która jest w pełni zdefiniowana w nagłówku korpusu; fragment tej definicji został przedstawiony poniżej[6]:

```
<classDecl>
 <taxonomy xml:id="taxonomy-NKJP-type">
 <!-- ... --->
   <category xml:id="typ_lit_proza">
    <desc xml:lang="pl">proza</desc>
    <desc xml:lang="en">prose</desc>
   </category>
   <category xml:id="typ_lit_poezja">
    <desc xml:lang="pl">poezja</desc>
    <desc xml:lang="en">poetry</desc>
   </category>
   <category xml:id="typ_lit_dramat">
    <desc xml:lang="pl">dramat</desc>
    <desc xml:lang="en">drama</desc>
   </category>
 <!-- ... --->
 </taxonomy>
</classDecl>
```

Ostatnia część elementu `<encodingDesc>` w nagłówku NKJP, o której chcielibyśmy wspomnieć, to `<nkjp:fsLib>`. Jak sugeruje prefiks `nkjp`, element ten nie jest częścią standardu TEI, lecz został zdefiniowany w ramach projektu NKJP. Było to konieczne, gdyż TEI – choć zawiera standard ISO reprezentacji struktur atrybutów w postaci XML (ISO:24610-1 2005) – nie pozwala na zdefiniowanie biblioteki takich struktur (ang. *feature structure library*) w nagłówku korpusu. Więcej o przydatności takiej definicji powiemy w p. 10.7. Tutaj zauważmy tylko, że obecność w reprezentacji takich elementów i atrybutów `nkjp:...` czyni opisywany w tym rozdziale schemat XML *rozszerzeniem* TEI (ang. *TEI Extension*), a nie *podzbiorem* standardu (ang. *TEI Conformance*, Burnard i Bauman 2008: p. 23.3).

[6] Tu i w przykładach poniższych zapis typu `<!- ... -->` sygnalizuje pominięty fragment znakowania.

Niemniej jednak, jak zobaczymy poniżej, rozszerzenia takie były w schemacie wykorzystywanym w NKJP wprowadzane tylko wtedy, gdy było to zupełnie niezbędne, więc są one bardzo nieliczne[7].

10.3.2. Nagłówek tekstu

Ogólna struktura nagłówków poszczególnych tekstów, tj. plików `header.xml`, jest podobna do struktury nagłówka korpusu. Element `<teiHeader>` składa się z trzech części: `<fileDesc>`, `<profileDesc>` i `<revisionDesc>`. Ostatnia z nich, `<revisionDesc>`, podobnie jak w wypadku nagłówka korpusu, zawiera ciąg elementów `<change>` opisujących modyfikacje któregokolwiek z plików reprezentujących dany tekst i jego znakowanie.

Zawartość elementu `<profileDesc>` różni się natomiast od odpowiadającego mu elementu w nagłówku korpusu, gdyż zawiera wyłącznie jeden element, `<textClass>`. W elemencie tym znajdują się informacje dotyczące klasyfikacji tekstu według taksonomii przyjętych w NKJP (zob. rozdz. 2). Na przykład zawartość elementu `<profileDesc>` dla powieści Manueli Gretkowskiej *Namiętnik* wygląda następująco:

```
<profileDesc>
 <textClass>
  <classCode scheme="#ukd">821.162.1-3</classCode>
  <keywords scheme="#bn">
   <list>
    <item>Opowiadanie polskie -- 20 w.</item>
   </list>
  </keywords>
  <catRef scheme="#taxonomy-NKJP-type"
   target="#typ_lit_proza"/>
  <catRef scheme="#taxonomy-NKJP-channel"
   target="#kanal_ksiazka"/>
 </textClass>
</profileDesc>
```

Mogą się tu znajdować odnośniki do czterech klasyfikacji: dwóch zewnętrznych w stosunku do NKJP (wspomniana powyżej Uniwersalna Klasyfikacja Dziesiętna oraz klasyfikacja Biblioteki Narodowej; zob. `#bn`) i dwóch wewnętrznych (typ tekstu: `#typ_lit_proza` oraz kanał: `#kanal_ksiazka`).

[7] W sumie zostały wprowadzone cztery nowe elementy (`<nkjp:fsLib>`, `<nkjp:file>`, `<nkjp:topic>`, `<nkjp:paren>`) oraz cztery nowe atrybuty (`@nkjp:subcorpus`, `@nkjp:nps`, `@nkjp:manual`, `@nkjp:rejected`). Ich znaczenie jest wyjaśnione w odpowiednich punktach niniejszego rozdziału.

Ostatnia część nagłówka tekstu, <fileDesc>, zawiera szereg różnorodnych informacji o samym tekście: jego tytuł w NKJP (np. „TEI P5 encoded version of "Namiętnik"”), informacje bibliograficzne o tekście źródłowym (tytuł, autor, wydawca, data wydania itp.), notkę tym, jak tekst został pozyskany do NKJP (np. „<note type="text_origin">IPI PAN Corpus</note>”), a jeżeli pochodzi z wcześniejszego korpusu, to także nagłówek tekstu w tym korpusie. W części tej znajduje się także element <publicationStmt>, np.:

```
<publicationStmt nkjp:subcorpus="balanced">
 <availability status="restricted">
  <p>For all NKJP purposes.</p>
 </availability>
</publicationStmt>
```

Atrybut @nkjp:subcorpus jest kolejnym przykładem rozszerzenia standardu TEI, wprowadzonym tutaj m.in. w celu formalnej reprezentacji informacji o tym, czy dany tekst wchodzi do podkorpusu zrównoważonego.

W wypadku transkrypcji danych mówionych nagłówek może także zawierać informację o osobie odpowiedzialnej za dokonanie transkrypcji (wewnątrz elementu <respStmt> w <fileDesc>), o pochodzeniu tekstu (informacje innego typu niż w wypadku tekstów pisanych), a także kolejny wprowadzony w projekcie element, <nkjp:topic>, opisujący temat konwersacji. Ponadto oprócz <textClass> w części <profileDesc> znajdują się także: element <langUsage> opisujący poziom formalności konwersacji, element <particDesc> zawierający informacje socjologiczne o uczestnikach nagrania, oraz <settingDesc> z informacją o czasie i miejscu nagrania. Poniższy przykład ilustruje niektóre z elementów typowych dla danych konwersacyjnych:

```
<langUsage>
 <language ident="pl-x-formal"/>
</langUsage>
<nkjp:topic xml:lang="pl">Rozmowa o immunitecie Zbigniewa Ziobro,
 sytuacji w Gruzji i reakcji Unii Europejskiej na nią.</nkjp:topic>
<particDesc>
 <!-- ... --->
 <person xml:id="sp2" role="speaker">
  <persName>Zbigniew Ziobro</persName>
  <sex value="1">male</sex>
  <education xml:lang="pl">wyższe</education>
  <age>40</age>
  <residence>unknown</residence>
 </person>
 <!-- ... --->
```

```
</particDesc>
<settingDesc>
 <setting>
  <name type="place">TVP Info</name>
  <name type="voivodship" xml:lang="pl">mazowieckie</name>
  <date type="recorded" when="2008-09-02"/>
 </setting>
</settingDesc>
```

10.4. Struktura tekstu

Dla każdego tekstu wchodzącego w skład NKJP podstawowym plikiem zawierającym ten tekst i znakowanie jego struktury jest `text_structure.xml`.

Ogólny schemat pliku `text_structure.xml` dla pojedynczego tekstu przedstawiony jest poniżej; element `<front>` (część początkowa tekstu) i `<back>` (część końcowa) są opcjonalne (i nigdy nie występują w transkrypcjach danych mówionych):

```
<teiCorpus
 xmlns:xi="http://www.w3.org/2001/XInclude"
 xmlns="http://www.tei-c.org/ns/1.0">
 <xi:include href="NKJP_header.xml"/>
 <TEI>
  <xi:include href="header.xml"/>
  <text xml:id="struct_text">
   <front><!--  część początkowa  --></front>
   <body><!--   część główna      --></body>
   <back><!--   część końcowa     --></back>
  </text>
 </TEI>
</teiCorpus>
```

W schemacie tym należy zwrócić uwagę przede wszystkim na to, że tekst jest reprezentowany jako element `<teiCorpus>` i że logicznie zawiera nie tylko nagłówek tekstu (`<xi:include href="header.xml"/>`), ale też nagłówek całego korpusu (`<xi:include href="NKJP_header.xml"/>`).

W wypadku elementów `<front>` i `<back>` dopuszczalna jest w NKJP pełna gama elementów przewidzianych w standardzie TEI P5. Typowa zawartość elmentu `<front>` zawiera informację o tytule, być może z wyróżnieniem podtytułu, np.:

```
<docTitle>
 <titlePart type="main">Pieśni nędzy i zagłady</titlePart>
 <titlePart type="sub">Twórczość Mordechaja Gebirtiga
```

```
  w Salonie Poezji</titlePart>
</docTitle>
```

Zawartość elementu <body> została natomiast w NKJP mocno ograniczona w stosunku do możliwości oferowanych przez TEI P5. W wypadku tekstów mówionych dozwolony jest tu jedynie ciąg elementów <u> (wypowiedź; ang. *utterance*), z opcjonalnymi elementami <incident> (wydarzenie) między nimi, np.:

```
<body xml:id="txt_body">
 <u who="#sp3" xml:id="u1">ale zostaw to w ogóle dajcie buziaka
  przepraszam was laski</u>
 <u trans="overlap" who="#sp1" xml:id="u2">no dałyśmy sobie buziaka
  no</u>
 <!-- ... -->
</body>
```

W wypadku tekstów pisanych głównymi cegiełkami elementu <body> są natomiast akapity <p> (ang. *paragraph*), elementy <ab> (ang. *anonymous block*, tj. fragmenty tekstu o rozmiarze z grubsza akapitów, ale bez gwarancji, że odpowiadają akapitom logicznym w tekście wejściowym) oraz <head> (dla tytułów rozdziałów itp.). Cegiełki te mogą być – za pomocą być może zagnieżdżonych elementów z odpowiednimi wartościami atrybutu @type – pogrupowane w rozdziały, podrozdziały itd., np.:

```
<body>
 <!-- ... -->
 <div type="chapter" n="1">
  <head>Rozdział 1 Skąd się biorą paradygmaty?</head>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
 </div>
 <div type="chapter" n="2">
  <head>Rozdział 2 Świat według Pszczółki Mai</head>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
  <p><!-- ... --></p>
 </div>
 <!-- ... -->
</body>
```

Zawartość elementów <u>, <p>, <ab> i <head> jest w zasadzie czystym tekstem. Jedyne elementy XML, które mogą w nim wystąpić to: <gap> (dla oznaczenia

miejsc w tekście, w których w tekście źródłowym występowały tabele, rysunki itp.), <hi> z obligatoryjnym atrybutem @rend określającym rodzaj zaznaczenia tekstu (np. pogrubienie czy kursywa; ten element występuje wyłącznie w tekstach pisanych), <lb> (element zawsze pusty, oznaczający łamanie wiersza w poezji itp.; tylko w tekstach pisanych) i dwa elementy wykorzystywane w tekstach mówionych do zasygnalizowania zdarzeń niewerbalnych: <vocal> (zjawiska głosowe ale nie leksykalne) i <incident> (inne zjawiska niewerbalne).

Powyższe ograniczenia dotyczące elementów <u>, <p> i <ab> spowodowane są faktem, że ich zawartość jest znakowana lingwistycznie i indeksowana na potrzeby przeszukiwarki korpusowej Poliqarp, a więc wszelkie dodatkowe znaczniki wewnątrz tych elementów jedynie utrudniłyby dalsze przetwarzanie tekstów. Formalnie rzecz ujmując, przetwarzane na dalszych etapach są elementy zdefiniowane następującym wyrażeniem XPath: //body//(p|ab|u), tj. wyżej wymienione trzy typy elementów dowolnie głęboko zagnieżdżone w elemencie <body>.

10.5. Segmentacja

Ogólna struktura każdego poziomu lingwistycznego jest taka sama i podobna do struktury text_structure.xml:

```
<teiCorpus
 xmlns:xi="http://www.w3.org/2001/XInclude"
 xmlns="http://www.tei-c.org/ns/1.0">
 <xi:include href="NKJP_header.xml"/>
 <TEI>
  <xi:include href="header.xml"/>
  <text>
   <body>
    <!-- ciąg elementów <p> -->
   </body>
  </text>
 </TEI>
</teiCorpus>
```

Każdy element <p> odpowiada jednemu elementowi <p>, <ab> lub <u> w elemencie <body> pliku text_structure.xml. Ta odpowiedniość jest formalnie oddana przez użycie zdefiniowanego w TEI atrybutu @corresp, którego wartością jest identyfikator odpowiedniego elementu w text_structure.xml. Na przykład następującemu wierszowi w text_structure.xml:

```
<u who="sp0" xml:id="u-1"> przegląd prasy</u>
```

odpowiada następujący fragment w `ann_segmentation.xml`:

```
<p corresp="text_structure.xml#u-1" xml:id="segm_u-1">
 <s xml:id="segm_u-1.1-s">
  <!-- przegląd -->
  <seg corresp="text_structure.xml#string-range(u-1,1,8)"
   xml:id="segm_u-1.1-seg"/>
  <!-- prasy -->
  <seg corresp="text_structure.xml#string-range(u-1,10,5)"
   xml:id="segm_u-1.2-seg"/>
 </s>
</p>
```

Inne znaczniki struktury tekstu nie są przenoszone na poziomy lingwistyczne.

Jak pokazuje powyższy przykład, akapity w pliku segmentacyjnym zawierają informacje o dwóch rodzajach segmentacji: na zdania (elementy `<s>`; tutaj tylko jedno) oraz na segmenty wyrazowe (elementy `<seg>`), przy czym te drugie są zdefiniowane jako ciągi znaków w odpowiednim elemencie pliku `text_structure.xml` (zob. „`corresp="text_structure.xml#string-range(u-1,1,8)"`” powyżej). Na przykład pierwszy segment wyrazowy w tym przykładzie to sekwencja ośmiu znaków, poczynając od znaku drugiego (znak pierwszy, tj. spacja przed słowem `przegląd`, miałby numer 0) w elemencie o identyfikatorze `u-1` w pliku `text_structure.xml`. Segment ten otrzymuje swój własny identyfikator, `segm_u-1.1-seg`, dzięki czemu możliwe są odniesienia do niego z wyższych poziomów znakowania lingwistycznego.

Opisywany poziom może zawierać także informację o niejednoznacznościach segmentacyjnych, np. o niejednoznaczności napisu *Kiedyś* (*Kiedyś to zrobiłem.* vs *Kiedyś to zrobił?*; por. p. 6.2.2 i p. 6.6.3). Aby adekwatnie takie niejednoznaczności reprezentować, trzeba było wprowadzić rozszerzenie standardu TEI, tj. element `<nkjp:paren>`, umożliwiający grupowanie segmentów wewnątrz (standardowego) elementu `<choice>`[8]:

```
<s xml:id="segm_p-91.1-s">
 <choice>
  <nkjp:paren>
   <!-- Kiedy -->
   <seg corresp="text_structure.xml#string-range(p-91,0,5)"
    nkjp:rejected="true" xml:id="segm_p-91.3182-seg"/>
   <!-- ś -->
   <seg corresp="text_structure.xml#string-range(p-91,5,1)"
    nkjp:nps="true" nkjp:rejected="true" xml:id="segm_p-91.3183-seg"/>
```

[8] Drugie użycie elementu `<nkjp:paren>` w poniższym przykładzie nie jest konieczne.

```
  </nkjp:paren>
  <nkjp:paren>
   <!-- Kiedyś -->
   <seg corresp="text_structure.xml#string-range(p-91,0,6)"
    xml:id="segm_p-91.3184-seg"/>
  </nkjp:paren>
 </choice>
 <!-- opisał -->
 <seg corresp="text_structure.xml#string-range(p-91,7,6)"
  xml:id="segm_p-91.3185-seg"/>
 <!-- em -->
 <seg corresp="text_structure.xml#string-range(p-91,13,2)"
  nkjp:nps="true" xml:id="segm_p-91.3186-seg"/>
 <!-- ... -->
</s>
```

W powyższym przykładzie, gdyby nie został użyty grupujący element `<nkjp:paren>`, zasygnalizowany by został wybór między trzema segmentami (*Kiedy*, *ś* i *Kiedyś*), a nie między ciągiem dwóch segmentów (*Kiedy* plus *ś*) i segmentem trzecim (*Kiedyś*).

Na potrzeby segmentacji zostały wprowadzone także dwa nowe atrybuty: `@nkjp:rejected`, służący do oznaczania wariantów uznanych za niepoprawne w danym kontekście (powyżej odrzucony jest wariant dwusegmentowy napisu *Kiedyś*), oraz `@nkjp:nps`, informujący o braku poprzedzającej spacji (ang. *no preceding space*) przed danym segmentem, tj. o połączonych segmentach, jak *opisał*|*em* i *Kiedy*|*ś* powyżej. Szczegółowe uzasadnienie dwóch z tych rozszerzeń (`<nkjp:paren>` i `@nkjp:nps`) znajduje się w artykule Bańskiego i Przepiórkowskiego (2009).

10.6. Sensy słów

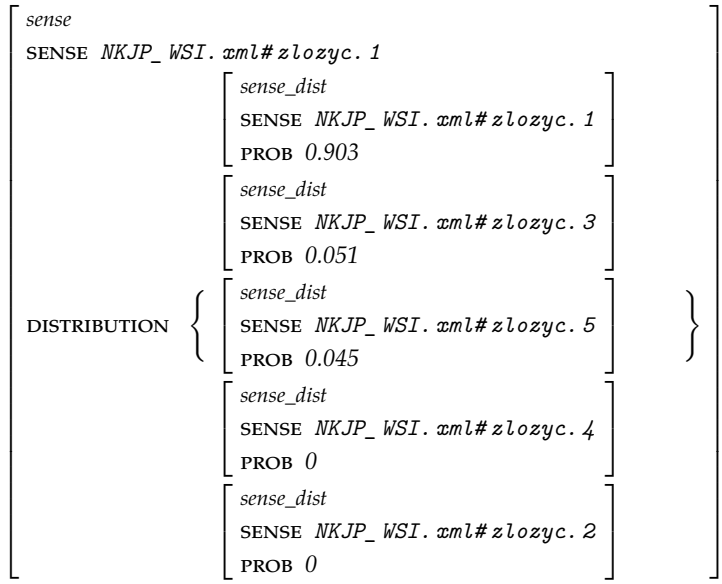
Ogólna struktura plików `ann_senses.xml` jest analogiczna do struktury plików `ann_segmentation.xml`, lecz znajdują się w nich reprezentacje jedynie tych segmentów, które są formami leksemów niejednoznacznych zdefiniowanych w słowniku `NKJP_WSI.xml`[9]; jeżeli w danym zdaniu czy akapicie takie formy nie wystąpiły, odpowiednie elementy `<s>` i `<p>` w pliku `ann_senses.xml` są puste. Dla każdej takiej formy, w `ann_senses.xml` obecny jest element `<seg>` zawierający XML-owe kodowanie struktury atrybutów typu *sense*. Struktury tego typu definiują dwa atrybuty: SENSE, wskazujący na znaczenie formy w danym kontekście,

[9] Lingwistyczne i implementacyjne zagadnienia związane z poziomem sensów słów zostały opisane w rozdziałach 7 i 12.

oraz – opcjonalnie – DISTRIBUTION, którego wartości określają dystrybucję prawdopodobieństw poszczególnych sensów danej formy według narzędzia, które dokonało automatycznego ujednoznacznienia sensów:

```
<s corresp="ann_segmentation.xml#segm_p-3.1-s"
   xml:id="senses_p-3.1-s">
 <seg corresp="ann_segmentation.xml#segm_p-3.123-seg"
  xml:id="senses_p-3.123-seg">
  <!-- złożyły [24,7] -->
  <fs type="sense">
   <f fVal="NKJP_WSI.xml#zlozyc.1" name="sense"/>
   <f name="distribution">
    <vColl>
     <fs type="sense_dist">
      <f fVal="0.903" name="prob"/>
      <f fVal="NKJP_WSI.xml#zlozyc.1" name="sense"/>
     </fs>
     <fs type="sense_dist">
      <f fVal="0.051" name="prob"/>
      <f fVal="NKJP_WSI.xml#zlozyc.3" name="sense"/>
     </fs>
     <fs type="sense_dist">
      <f fVal="0.045" name="prob"/>
      <f fVal="NKJP_WSI.xml#zlozyc.5" name="sense"/>
     </fs>
     <fs type="sense_dist">
      <f fVal="0" name="prob"/>
      <f fVal="NKJP_WSI.xml#zlozyc.4" name="sense"/>
     </fs>
     <fs type="sense_dist">
      <f fVal="0" name="prob"/>
      <f fVal="NKJP_WSI.xml#zlozyc.2" name="sense"/>
     </fs>
    </vColl>
   </f>
  </fs>
 </seg>
</s>
```

Zauważmy, że elementy `<fs>` (ang. *feature structure*) itp. są zdefiniowane we wspomnianym powyżej standardzie ISO będącym częścią TEI P5. Powyższa struktura może być graficznie przedstawiona jak w (10.1) na następnej stronie. A zatem forma *złożyły* została w tym kontekście ujednoznaczniona do znaczenia `zlozyc.1` (*połączyć części w całość*), a za zupełnie nieprawdopodobne zostało uznane m.in. znaczenie `zlozyc.2` (*dać pewną sumę pieniędzy, złożyć się na coś*).

(10.1)


10.7. Morfoskładnia

Każdy segment wyrazowy wyróżniony w `ann_segmentation.xml`, który nie został odrzucony (w sensie atrybutu `@nkjp:rejected`), ma przypisaną interpretację morfosyntaktyczną w pliku `ann_morphosyntax.xml`[10]. Podobnie jak w pliku `ann_senses.xml`, każdy segment jest reprezentowany jako element `<seg>` zawierający strukturę atrybutów, lecz w wypadku poziomu morfosyntaktycznego struktura ta jest znacznie bardziej rozbudowana:

```
<seg corresp="ann_segmentation.xml#segm_p-1.1-seg"
 xml:id="morph_p-1.1-seg">
 <fs type="morph">
  <f name="orth">
   <string>Wstęp</string>
  </f>
  <!-- Wstęp [0,5] -->
  <f name="interps">
   <fs type="lex">
    <f name="base">
     <string>wstęp</string>
    </f>
    <f name="ctag">
     <symbol value="subst"/>
```

[10] Por. rozdziały 6 i 11.

```
    </f>
    <f name="msd">
     <vAlt>
      <symbol value="sg:nom:m3" xml:id="morph_p-1.1-seg_0-msd"/>
      <symbol value="sg:acc:m3" xml:id="morph_p-1.1-seg_1-msd"/>
     </vAlt>
    </f>
   </fs>
  </f>
  <f name="disamb">
   <fs feats="#pantera" type="tool_report">
    <f fVal="#morph_p-1.1-seg_0-msd" name="choice"/>
    <f name="interpretation">
     <string>wstęp:subst:sg:nom:m3</string>
    </f>
   </fs>
  </f>
 </fs>
</seg>
```

Zauważmy, że struktura atrybutów typu *tool_report*, będąca wartością atrybutu DISAMB, zawiera specyfikację `feats="#pantera"`. Jest to odniesienie do następującej struktury atrybutów zdefiniowanej w nagłówku korpusu, we wspomnianym powyżej elemencie `<nkjp:fsLib>`:

```
<nkjp:fsLib>
 <fLib n="tools">
  <f xml:id="pantera" name="tool">
   <string>PANTERA Tagger (April 2011)</string>
  </f>
 </fLib>
</nkjp:fsLib>
```

A zatem bardziej czytelna reprezentacja pełnej struktury wygląda następująco:

(10.2)
$$\begin{bmatrix} \textit{morph} \\ \text{ORTH} \ \textit{Wstęp} \\ \text{INTERPS} \begin{bmatrix} \textit{lex} \\ \text{BASE} \ \textit{wstęp} \\ \text{CTAG} \ \textit{subst} \\ \text{MSD} \ \boxed{0}\textit{sg:nom:m3} \mid \textit{sg:acc:m3} \end{bmatrix} \\ \text{DISAMB} \begin{bmatrix} \textit{tool\_report} \\ \text{TOOL} \ \textit{PANTERA Tagger (April 2011)} \\ \text{CHOICE} \ \boxed{0} \\ \text{INTERPRETATION} \ \textit{wstęp:subst:sg:nom:m3} \end{bmatrix} \end{bmatrix}$$

Mówi ona tyle, że dany segment to *Wstęp*, że ma dwie interpretacje jako formy rzeczownika wstęp oraz że w danym kontekście narzędzie PANTERA (zob. rozdz. 11) wybrało pierwszą z nich, tj. interpretację mianownikową (zob. `@choice`; dwa wystąpienia zmiennej 0 oznaczają odwołanie do tej samej wartości, tj. tutaj do sg:nom:m3)[11]. W wypadku słów o interpretacjach różniących się klasą gramatyczną lub formą podstawową, wartością atrybutu INTERPS jest lista struktur typu *lex*.

10.8. Poziomy składniowe

Przez poziomy składniowe znakowania lingwistycznego będziemy tutaj rozumieć poziomy słów składniowych, grup składniowych oraz jednostek nazewniczych[12]. Od poziomów opisanych w poprzednich dwóch punktach odróżnia je to, że jednostki w nich wyodrębnione zawierają mniejsze jednostki. Oznacza to, że elementy `<seg>` mają inną interpretację niż na poziomach omówionych powyżej oraz że – oprócz struktury atrybutów opisującej daną jednostkę – zawierają ciąg elementów `<ptr>`, wskazujących na składniki bezpośrednie danej jednostki.

```
<p xml:id="p-1" corresp="ann_morphosyntax.xml#p-1">
 <s xml:id="p-1.1-s" corresp="ann_morphosyntax.xml#p-1.1-s"/>
</p>
```

10.8.1. Słowa składniowe

W najprostszym wypadku jedynym składnikiem słowa składniowego jest segment, jak w poniższym przykładzie:

```
<seg xml:id="words_1.1-s_11">
 <fs type="word">
  <f name="orth">
   <string>jest</string>
  </f>
  <f name="interps">
   <fs type="lex">
    <f name="base">
     <string>być</string>
    </f>
    <f name="ctag">
     <symbol value="Verbfin"/>
    </f>
```

[11] Atrybut `@interpretation` jest redundanty w stosunku do reszty struktury, lecz jego obecność ułatwia przetwarzanie na dalszych etapach.

[12] Por. rozdziały 8 i 9, a także 13.

```
    <f name="msd">
     <symbol value="sg:ter:pres:ind:imperf:nrefl:aff"/>
    </f>
   </fs>
  </f>
 </fs>
 <ptr target="ann_morphosyntax.xml#morph_1.1.9-seg"/>
</seg>
```

Jedyną funkcją takiego słowa składniowego jest „opakowanie" segmentu zinterpretowanego w pliku `ann_morphosyntax.xml` i nadanie mu nowego znacznika morfosyntaktycznego: zamiast fin:sg:ter:imperf jest to Verbfin:sg:ter:pres:ind:imperf:nrefl:aff. Poza brakiem atrybutu DISAMB[13], struktura atrybutów dla słowa składniowego jest analogiczna do struktury dla anotacji morfosyntaktycznej segmentu.

Słowa składniowe mogą składać się z kilku segmentów zinterpretowanych w `ann_morphosyntax.xml` lub – co nie zostało w praktyce wykorzystane w NKJP – z mniejszych słów składniowych zdefiniowanych w bieżącym pliku; w takich wypadkach reprezentujący je element `<seg>` zawiera kilka elementów `<ptr>`. Zauważmy, że reprezentacja składników bezpośrednich za pomocą odnośników typu `<ptr>` pozwala na reprezentację nieciągłości i częściowego pokrywania się słów składniowych. Ciekawym przypadkiem jest tu haplologia *się* (Kupść 1999), jak w zdaniu *Bał się zaśmiać*, w którym występują dwa czasowniki zwrotne (a więc słowa składniowe dwusegmentowe), a tylko jedno *się*. Schematyczna reprezentacja tego zdania na poziomie słów składniowych przedstawiona jest poniżej, przy czym dla zwiększenia czytelności, struktury atrybutów opisujące słowa podane są wyłącznie w postaci graficznej.

```
<s>
 <seg>
  <fs><!-- 1 --></fs>  <!-- (zob. poniżej) -->
  <ptr target="ann_morphosyntax.xml#seg17"/>  <!-- Bał -->
  <ptr target="ann_morphosyntax.xml#seg18"/>  <!-- się -->
 </seg>
 <seg>
  <fs><!-- 2 --></fs>  <!-- (zob. poniżej) -->
  <ptr target="ann_morphosyntax.xml#seg18"/>  <!-- się -->
  <ptr target="ann_morphosyntax.xml#seg19"/>  <!-- zaśmiać -->
 </seg>
</s>
```

[13] Zakładamy, że na podstawie jednoznacznie zinterpretowanych morfosyntaktycznie segmentów można nadać jednoznaczną interpretację słowom składniowym. Być może w przyszłości zajdzie potrzeba zrezygnowania z tego założenia i dodania atrybutu DISAMB.

(10.3) $\boxed{1} =$
$$\begin{bmatrix} \textit{word} \\ \textsc{orth}\ \textit{Bał się} \\ \textsc{interps}\ \begin{bmatrix} \textit{lex} \\ \textsc{base}\ \textit{bać się} \\ \textsc{ctag}\ \textit{Verbfin} \\ \textsc{msd}\ \textit{sg:ter:past:ind:imperf:refl:aff:m1} \end{bmatrix} \end{bmatrix}$$

(10.4) $\boxed{2} =$
$$\begin{bmatrix} \textit{word} \\ \textsc{orth}\ \textit{się zaśmiać} \\ \textsc{interps}\ \begin{bmatrix} \textit{lex} \\ \textsc{base}\ \textit{zaśmiać się} \\ \textsc{ctag}\ \textit{Inf} \\ \textsc{msd}\ \textit{perf:refl:aff} \end{bmatrix} \end{bmatrix}$$

10.8.2. Grupy składniowe

Grupy składniowe odnoszą się do poziomu słów składniowych. Jak pokazuje poniższy przykład, struktura atrybutów opisująca grupę składniową jest szczególnie prosta, gdyż zawiera tylko dwa atrybuty: ORTH (postać ortograficzna grupy) i TYPE (typ grupy):

```
<seg xml:id="groups_1.9-s_gae">
 <fs type="group">
  <f name="orth">
   <string>Tego dnia</string>
  </f>
  <f name="type">
   <symbol value="NG"/>
  </f>
 </fs>
 <ptr type="nonhead" target="ann_words.xml#words_1.9-s_sa6"/>
 <ptr type="head" target="ann_words.xml#words_1.9-s_sab"/>
</seg>
```

Dodatkową różnicą w stosunku do poziomu słów składniowych jest nowy atrybut elementów `<ptr>`, a mianowicie `@type`, informujący o tym, czy dany składnik bezpośredni jest składniowym (`type="synh"`) lub semantycznym (`type="semh"`) elementem głównym grupy; jeżeli jest on elementem głównym jednocześnie obu typów, wartością atrybutu `@type` jest `head`, a jeżeli nie jest żadnym elementem głównym, wartością tą jest `nonhead`, jak w powyższym przykładzie.

10.8.3. Jednostki nazewnicze

Choć docelowo jednostki nazewnicze powinny zapewne odnosić się do poziomu słów składniowych, w obecnej wersji korpusu odnoszą się one do poziomu morfoskładniowego.

Jak zostało to powiedziane w rozdz. 9, jednostki nazewnicze mogą być zagnieżdżone, na przykład jednostka oznaczająca pełne imię i nazwisko może zawierać dwie mniejsze jednostki odpowiadające imieniu i nazwisku:

```
<seg xml:id="p-6.1-s_n4">
 <fs type="named">
  <f name="type">
   <symbol value="persName"/>
  </f>
  <f name="orth">
   <string>"Mirosława Formelę"<string/>
  </f>
 </fs>
 <ptr target="p-6.1-s_n5"/>
 <ptr target="p-6.1-s_n6"/>
</seg>
<seg xml:id="p-6.1-s_n5">
 <fs type="named">
  <f name="type">
   <symbol value="persName"/>
  </f>
  <f name="subtype">
   <symbol value="forename"/>
  </f>
  <f name="orth">
   <string>"Mirosława"<string/>
  </f>
 </fs>
 <ptr target="ann_morphosyntax.xml#morph_p-6.206-seg"/>
</seg>
<seg xml:id="p-6.1-s_n6">
 <fs type="named">
  <f name="type">
   <symbol value="persName"/>
  </f>
  <f name="subtype">
   <symbol value="surname"/>
  </f>
  <f name="orth">
   <string>"Formelę"<string/>
  </f>
 </fs>
 <ptr target="ann_morphosyntax.xml#morph_p-6.207-seg"/>
</seg>
```

W powyższym przykładzie zostały wyróżnione dwie jednostki jednosegmentowe (*Mirosława* i *Formelę*) oraz zawierająca je (w sensie odnośników <ptr>) jednostka złożona *Mirosława Formelę*. Zwróćmy uwagę na to, że każda z tych jednostek jest opisana strukturą atrybutów zawierającą obligatoryjnie atrybuty ORTH i TYPE, oraz – opcjonalnie – SUBTYPE.

10.9. Korpus ręcznie znakowany

Opisany w rozdz. 5 ręcznie znakowany podkorpus wielkości ok. miliona słów jest znakowany podobnie jak cały NKJP. Niniejszy punkt opisuje drobne różnice, wynikające ze specyfiki znakowania ręcznego.

10.9.1. Metadane

Struktura nagłówków korpusu milionowego i poszczególnych próbek wchodzących w jego skład jest taka sama jak w wypadku pełnego NKJP, tyle że w nagłówkach próbek znajduje się informacja o liczbie słów w danym pliku, np:

```
<extent nkjp:file="text.xml">
 <num type="word" value="10005"/>
</extent>
```

Ponieważ każdy tekst w korpusie jest reprezentowany w postaci kilku plików XML, konieczne stało się dodanie do elementu <extent> atrybutu @nkjp:file; wskazuje on na plik, w którym faktycznie znajdują się liczone słowa.

Oczywiście same informacje wypełniające tę samą strukturę nagłówków pełnego korpusu NKJP i jego podkorpusu w wielu miejscach się różnią. Na przykład w nagłówku NKJP element <nkjp:fsLib> został wykorzystany do definicji opisu tagera morfosyntaktycznego, podczas gdy w nagłówku podkorpusu 1M – do opisu dwóch wersji Anotatorni (por. p. 6.6.1):

```
<nkjp:fsLib>
 <fLib n="tools">
  <f xml:id="an8003" name="tool">
   <string>Anotatornia NKJP on port 8003</string>
  </f>
  <f xml:id="an8004" name="tool">
   <string>Anotatornia NKJP on port 8004</string>
  </f>
 </fLib>
</nkjp:fsLib>
```

10.9.2. Struktura tekstu

Jak to wynika już z przykładu w poprzednim punkcie, pliki zawierające tekst w korpusie znakowanym ręcznie nazywają się `text.xml`, a nie `text_structure.xml`. Inna nazwa sygnalizuje, że faktycznie brak w tych plikach informacji strukturalnych: brak w nich elementów `<front>` i `<back>`, zaś `<body>` to ciąg elementów `<div>` zawierających fragmenty tekstu ujęte w elementy `<ab>`, a więc bez gwarancji, że są to pełne akapity logiczne.

10.9.3. Segmentacja

Struktura poziomu segmentacyjnego w podkorpusie znakowanym ręcznie jest identyczna z tą dla całego NKJP, z oczywistą różnicą wynikającą z poprzedniego punktu: segmenty są definiowane przez odniesienie się do odpowiednich elementów `<ab>` w pliku `text.xml`, np.:

```
<seg corresp="text.xml#string-range(txt_12.1-ab,164,7)"
 xml:id="segm_12.25-seg"/>
<!-- wynosi -->
<seg corresp="text.xml#string-range(txt_12.1-ab,172,6)"
 xml:id="segm_12.26-seg"/>
<choice>
 <!-- 12Mbps -->
 <seg corresp="text.xml#string-range(txt_12.1-ab,179,6)"
  nkjp:rejected="true" xml:id="segm_12.27-seg"/>
 <nkjp:paren>
  <!-- 12 -->
  <seg corresp="text.xml#string-range(txt_12.1-ab,179,2)"
   xml:id="segm_9927.693133-seg"/>
  <!-- Mbps -->
  <seg corresp="text.xml#string-range(txt_12.1-ab,181,4)"
   nkjp:nps="true" xml:id="segm_9927.693134-seg"/>
 </nkjp:paren>
</choice>
```

10.9.4. Sensy słów

Struktura plików `ann_senses.xml` w podkorpusie 1M jest identyczna jak w całym korpusie NKJP.

10.9.5. Morfoskładnia

Na poziomie morfoskładniowym ręcznie znakowanego podkorpusu NKJP zostało wykorzystane ostatnie z kilku rozszerzeń standardu TEI, a mianowicie atrybut `@nkjp:manual`. Jest on wykorzystany do oznaczenia tych interpretacji, które zostały w procesie znakowania wprowadzone ręcznie, gdyż nie były znane analizatorowi morfosyntaktycznemu. Użycie tego atrybutu ilustruje poniższy przykład, w którym do istniejących przymiotnikowych interpretacji segmentu *Lotną* została dodana ta właściwa w danym kontekście, tj. rzeczownikowa[14]:

```
<seg corresp="ann_segmentation.xml#segm_2.49-seg"
 xml:id="morph_2.49-seg">
 <fs type="morph">
  <f name="orth">
   <string>Lotną</string>
  </f>
  <f name="interps">
   <fs type="lex" xml:id="morph_2.49.1-lex">
    <f name="base">
     <string>lotny</string>
    </f>
    <f name="ctag">
     <symbol value="adj"/>
    </f>
    <f name="msd">
     <vAlt>
      <symbol value="sg:acc:f:pos" xml:id="morph_2.49.1.1-msd"/>
      <symbol value="sg:inst:f:pos" xml:id="morph_2.49.1.2-msd"/>
     </vAlt>
    </f>
   </fs>
   <fs type="lex" xml:id="morph_2.49.2-lex">
    <f name="base">
     <string>Lotna</string>
    </f>
    <f name="ctag">
     <symbol value="subst"/>
    </f>
    <f name="msd">
     <symbol nkjp:manual="true" value="sg:acc:f"
      xml:id="morph_2.49.2.1-msd"/>
    </f>
```

[14] Mowa oczywiście o imieniu konia ze znanego opowiadania Wojciecha Żukrowskiego sfilmowanego przez Andrzeja Wajdę.

```
   </fs>
  </f>
  <f name="disamb">
   <fs feats="#an8003" type="tool_report">
    <f fVal="#morph_2.49.2.1-msd" name="choice"/>
    <f name="interpretation">
     <string>Lotna:subst:sg:acc:f</string>
    </f>
   </fs>
  </f>
 </fs>
</seg>
```

10.9.6. Poziomy składniowe

Spośród trzech poziomów składniowych jedynie na poziomie jednostek nazewniczych struktura w korpusie znakowanym ręcznie stanowi pewne rozszerzenie w stosunku do pełnego NKJP: oprócz formy ortograficznej i typu (ew. także podtypu) danej jednostki, obecna jest także informacja o jej formie podstawowej oraz o stopniu pewności anotatora znakującego daną nazwę własną co do jego decyzji. Rozszerzenia te ilustruje poniższy przykład:

```
<seg xml:id="named_2.7-s_n4">
 <fs type="named">
  <f name="type">
   <symbol value="persName"/>
  </f>
  <f name="orth">
   <string>Hienadzia Karpienki</string>
  </f>
  <f name="base">
   <string>Hienadź Karpienko</string>
  </f>
  <f name="certainty">
   <symbol value="high"/>
  </f>
 </fs>
 <ptr target="named_2.7-s_n2"/>
 <ptr target="named_2.7-s_n3"/>
</seg>
<seg xml:id="named_2.7-s_n2">
 <fs type="named">
  <f name="type">
   <symbol value="persName"/>
```

```
  </f>
  <f name="subtype">
   <symbol value="forename"/>
  </f>
  <f name="orth">
   <string>Hienadzia</string>
  </f>
  <f name="base">
   <string>Hienadź</string>
  </f>
  <f name="certainty">
   <symbol value="high"/>
  </f>
 </fs>
 <ptr target="ann_morphosyntax.xml#morph_2.5-seg"/>
</seg>
<seg xml:id="named_2.7-s_n3">
 <fs type="named">
  <f name="type">
   <symbol value="persName"/>
  </f>
  <f name="subtype">
   <symbol value="surname"/>
  </f>
  <f name="orth">
   <string>Karpienki</string>
  </f>
  <f name="base">
   <string>Karpienko</string>
  </f>
  <f name="certainty">
   <symbol value="high"/>
  </f>
 </fs>
 <ptr target="ann_morphosyntax.xml#morph_2.6-seg"/>
</seg>
```

W wypadku gdy wartością atrybutu CERTAINTY jest *medium* lub *low*, pojawia się dodatkowy atrybut, COMMENT, zawierający uzasadnienie wątpliwości anotatora. Możliwość tę ilustrujemy przykładem, w którym występuje jednostka nazewnicza derywowana od leksemu ROSJA:

```
<seg xml:id="named_2.41-s_n1">
 <fs type="named">
  <f name="derived">
```

```
   <fs type="derivation">
    <f name="derivType">
     <symbol value="relAdj"/>
    </f>
    <f name="derivedFrom">
     <string>Rosja</string>
    </f>
   </fs>
  </f>
  <f name="type">
   <symbol value="placeName"/>
  </f>
  <f name="subtype">
   <symbol value="country"/>
  </f>
  <f name="orth">
   <string>ruski</string>
  </f>
  <f name="base">
   <string>ruski</string>
  </f>
  <f name="certainty">
   <symbol value="medium"/>
  </f>
  <f name="comment">
   <string>forma z nacechowaniem pejoratywnym</string>
  </f>
 </fs>
 <ptr target="ann_morphosyntax.xml#morph_2.39-seg"/>
</seg>
```

10.10. Podsumowanie

W niniejszym rozdziale przedstawiliśmy schemat znakowania informacji strukturalnych, bibliograficznych i lingwistycznych w Narodowym Korpusie Języka Polskiego. Schemat ten tylko nieznacznie wychodzi poza propozycje Text Encoding Initiative; jego główna wartość polega raczej na wybraniu spośród mechanizmów oferowanych przez TEI tych, które mogą posłużyć do spójnego znakowania korpusów językowych zawierających wiele poziomów anotacji lingwistycznej. Dosyć permisywne definicje poszczególnych poziomów znakowania dostępne są jako schematy ODD (ang. *One Document Does it all*) i Relax NG na stronie `http://nlp.ipipan.waw.pl/TEI4NKJP/`, na której znaleźć można także szersze przykłady poszczególnych poziomów i wybrane artykuły.

Część IV

Narzędzia i podprojekty

11

Tager morfosyntaktyczny PANTERA

Szymon Acedański

11.1. O narzędziu

Rozdział ten opisuje tager morfosyntaktyczny opracowany na potrzeby projektu NKJP. Korpus ten jest anotowany na wielu poziomach, a PANTERA odpowiada za ujednoznacznianie opisu morfosyntaktycznego tekstów pełnego korpusu. Technicznie przetwarzanie wykonywane przez PANTERĘ składa się z kilku etapów:

1. wykonanie podziału tekstu na zdania, przy użyciu narzędzia Segment (Miłkowski i Lipski 2011);
2. wykonanie analizy morfoskładniowej za pomocą programu Morfeusz SGJP (Woliński 2006); w przypadku słów nieznanych przez Morfeusza używany jest moduł Odgadywacz, który jest częścią tagera TaKIPI (Piasecki 2007);
3. wykonanie dezambiguacji segmentacji w oparciu o proste reguły leksykalne;
4. dezambiguacja opisu morfosyntaktycznego.

Dezambiguator oparty jest na algorytmie opisanym w pracy Erica Brilla (1992), z licznymi modyfikacjami dostosowującymi go do charakterystyki języków fleksyjnych, jak również z możliwością pracy równoległej. Dzięki nim udało się zwiększyć skuteczność tagowania do 92,23% – poprawa o 1,17% w porównaniu z najlepszym dotychczasowym wynikiem, opisanym w pracy Piaseckiego (2006) – i uzyskać zadowalającą szybkość tagowania.

Algorytm Brilla polega na automatycznym wygenerowaniu transformacji na podstawie danych treningowych, według ustalonych wcześniej szablonów. Transformacje te, zastosowane następnie do tekstu początkowo oznaczonego przy pomocy prostego tagera (np. unigramowego), znacząco zwiększają jakość

tagowania. Stąd też nazwa PANTERA jest akronimem pochodzącym od frazy „Polskiej Akademii Nauk Tager Ekstrahujący Reguły Automatycznie".

11.2. Algorytm

Na początku przedstawimy, jak działa oryginalny algorytm podany przez Brilla, przeznaczony dla języka angielskiego (mającego niewielki tagset). Załóżmy, że mamy dane trzy korpusy ręcznie anotowane: korpus treningowy, mniejszy korpus poprawek oraz korpus testowy (używany tylko do ewaluacji). Zakładamy także, że w tych korpusach każdemu segmentowi przypisany jest dokładnie jeden prawidłowy tag (ozn. t_i dla i-tego segmentu). Potrzebny będzie również dowolny prosty tager (zwany tagerem wstępnym), np. unigramowy[1].

1. Jeśli tager wstępny jest statystyczny, to najpierw używamy korpusu treningowego do jego wytrenowania.
2. Tagujemy korpus poprawek przy użyciu wytrenowanego tagera wstępnego, cały czas pamiętając tagi wzorcowe (przypisane ręcznie). Jakość tego tagowania zapewne będzie niezbyt zgodna z oczekiwanymi poprawnymi wartościami przypisanymi przez anotatorów. Za pomocą algorytmu Brilla będziemy chcieli wygenerować ciąg transformacji, których zastosowanie poprawi jak najwięcej błędów.
3. Transformacje są generowane na podstawie *szablonów transformacji*. Brill używa w swojej pracy następujących szablonów:
   a) $t_i := A$ jeśli $t_i = B \wedge \exists_{o \in O_1} t_{i+o} = C$,
   „Zmień tag z B na A, jeśli jeden z okolicznych segmentów (otoczenie O_1) ma tag C";
   b) $t_i := A$ jeśli $t_i = B \wedge \forall_{o \in O_2} t_{i+o} = C_o$,
   „Zmień tag z B na A, jeśli wszystkie sąsiednie segmenty (w otoczeniu O_2) mają tag C";
   c) $t_i := A$ jeśli $t_i = B$ i i-ty segment rozpoczyna się od wielkiej litery,
   „Zmień tag z B na A, jeśli bieżący segment rozpoczyna się od wielkiej litery";
   d) $t_i := A$ jeśli $t_i = B$ oraz $(i-1)$-szy segment rozpoczyna się od wielkiej litery,
   „Zmień tag z B na A, jeśli poprzedni segment rozpoczyna się od wielkiej litery";

[1] Tager unigramowy to taki, który każdemu segmentowi przypisuje interpretację najczęściej występującą w danych treningowych dla tej samej formy ortograficznej.

gdzie:
- $O_1 \in \{\{1\},\{-1\},\{2\},\{-2\},\{1,2\},\{-1,-2\},\{1,2,3\},$ $\{-1,-2,-3\}\}$,
- $O_2 \in \{\{-2,-1\},\{-1,1\},\{1,2\}\}$,
- zmienne O_1 oraz O_2 określają, w jakim otoczeniu rozpatrywanego segmentu szukamy konkretnych tagów (np. zbiór $\{-1,-2\}$ oznacza „jeden z dwóch poprzedzających segmentów"),
- A, B, C, C_o – dowolne tagi.

Rozważając wszystkie możliwe kombinacje zmiennych (A, B, C, C_o, O_1, O_2), tager w sposób automatyczny wybierze takie ich wartości, z których powstają transformacje poprawiające dużą liczbę błędów.

4. Dla każdej możliwej transformacji r wygenerowanej z powyższych szablonów liczymy dwie statystyki:
   a) $good(r)$ – liczba miejsc w korpusie poprawek, w którym reguła ma zastosowanie oraz zmienia tag z niepoprawnego na poprawny,
   b) $bad(r)$ – liczba miejsc w korpusie poprawek, w którym reguła ma zastosowanie oraz zmienia tag z poprawnego na niepoprawny.
5. Znajdujemy transformację r_b o największej wartości $good(r) - bad(r)$. Zapamiętujemy ją i stosujemy do korpusu poprawek. Jeśli wciąż są w nim błędy, powtarzamy poszukiwanie reguły aż do uzyskania żądanej liczby reguł bądź do wyczerpania reguł o wartościach $good(r) - bad(r)$ większych od ustalonego progu.

Po zakończeniu trenowania tagowanie dowolnego tekstu odbywa się dwuetapowo:

1. Najpierw tekst jest tagowany przy użyciu tego samego tagera wstępnego, co użyty podczas trenowania.
2. Aplikowane są kolejno wygenerowane transformacje.

11.3. Adaptacja algorytmu Brilla dla języków fleksyjnych

Opisany powyżej algorytm nie nadaje się do bezpośredniego zastosowania do języka polskiego, przede wszystkim z powodu istotnie innego tagsetu. Tagi dla polskiego mają wewnętrzną strukturę – wartości poszczególnych kategorii gramatycznych czasem są od siebie kompletnie niezależne. Dlatego też pierwotne szablony Brilla, które biorą pod uwagę jedynie całość tagów, nie mogą być tutaj skutecznie zastosowane. Duża liczba możliwych tagów w języku polskim jest także powodem dużej złożoności obliczeniowej podczas trenowania.

Dlatego też zastosowano szereg modyfikacji oryginalnego algorytmu:

1. Tagowanie wieloprzebiegowe – w każdym z przebiegów odbywa się tagowanie jedynie dla wybranych kategorii gramatycznych.
2. Uogólnione szablony transformacji – pozwalają na uwzględnienie wartości poszczególnych kategorii gramatycznych, a nie tylko całych tagów,
3. Leksykalne szablony transformacji – pozwalają na uwzględnienie konkretnych końcówek.
4. Uproszczona implementacja algorytmu FastTBL (Ngai i Florian 2001) oraz zrównoleglenie silnika tagowania dla uzyskania praktycznych szybkości przetwarzania.

11.3.1. Tagowanie wieloprzebiegowe

Metoda ta była z powodzeniem stosowana do języka czeskiego (Hajič 1997). Eksperymenty wykonane w pierwszej fazie powstawania PANTERY potwierdziły jej skuteczność także w wypadku języka polskiego (Acedański i Gołuchowski 2009). Polega ona na podziale tagowania na fazy. W pierwszej fazie tager decyduje o klasie gramatycznej każdego segmentu oraz wartościach tylko niektórych kategorii gramatycznych (w projekcie NKJP będą to przypadek oraz osoba). W kolejnej fazie ustalane są wartości pozostałych kategorii, bez możliwości zmiany decyzji podjętych w pierwszej fazie.

Dzięki takiemu podejściu zapobiegamy powstawaniu zbyt licznych transformacji z mało ogólnymi predykatami. W ten sposób zmniejszamy szansę przeuczenia się algorytmu.

11.3.2. Uogólnione szablony transformacji

W oryginalnym algorytmie Brilla, opisanym w p. 11.2, wszystkie szablony transformacji miały postać:

> Zmień t_i na A, jeżeli jest B, pod warunkiem, że …

Tworząc nowy tager, postanowiliśmy uogólnić te szablony, pozwalając wykonywać inne operacje niż tylko zmiana całego tagu z konkretnie ustalonego na inny ustalony. Na przykład zależy nam na tym, żeby tager mógł wygenerować taką transformację:

> Jeśli bieżącym segmentem jest przymiotnik ($t_i|_{\text{POS}} = adj$), a kolejnym rzeczownik ($t_{i+1}|_{\text{POS}} = subst$), uzgodnij ich przypadki ($t_i|_{\text{CASE}} := t_{i+1}|_{\text{CASE}}$).

A zatem szablon uogólniony (zwany odtąd po prostu szablonem) składa się z:

1. *predykatu* wyznaczającego warunki, jakie musi spełnić segment oraz jego kontekst, aby transformacja mogła zostać zastosowana;
2. *akcji* wykonywanej, jeśli predykat jest spełniony w danym miejscu.

W powyższym przykładzie akcją jest „$t_i|_{\text{CASE}} := t_{i+1}|_{\text{CASE}}$", natomiast predykatem „$(t_i|_{\text{POS}} = adj) \wedge (t_{i+1}|_{\text{POS}} = subst)$".

Jednocześnie przyjmujemy, że nie wykonujemy takich transformacji, które przypisywałyby danemu segmentowi interpretację spoza zbioru możliwych interpretacji zwróconych przez analizator morfosyntaktyczny. Innymi słowy traktujemy każdy predykat tak, jakby był rozszerzony o sprawdzenie, czy po wykonaniu akcji nowo przypisany tag jest wciąż dopuszczalny.

Podczas trenowania tagera NKJP wykorzystano następujące szablony predykatów:

1. $t_i = T \wedge \exists_{o \in O} t_{i+o} = U$,
   „Bieżący segment ma tag T, a jeden z sąsiednich (otoczenie O) ma tag U";
2. $t_i = T \wedge \forall_{o \in O} t_{i+o} = U_o$,
   „Bieżący segment ma tag T, a otoczenie O ma konkretne tagi (U_o)";
3. $t_i = T \wedge \exists_{o \in O_1} t^{P1}_{i+o} = U'$
   (tylko w drugiej fazie; t^{P1}_i oznacza tag i-tego segmentu z fazy pierwszej),
   „Bieżący segment ma tag T, a jeden z sąsiednich (otoczenie O) miał tag U' z poprzedniej fazy";
4. $t_i|_{\text{POS}} = P \wedge t^p_i|_C = X \wedge \exists_{o \in O} \left(t^p_{i+o}|_{\text{POS}} = Q \wedge t^p_{i+o}|_C = Y\right)$,
   „Bieżący segment ma część mowy P, kategoria gramatyczna C przyjmuje wartość X, a jeden z sąsiednich (otoczenie O) segmentów ma część mowy Q, a wartością jego kategorii C jest Y";
5. $t_i|_{\text{POS}} = P \wedge t_i|_C = X \wedge \forall_{o \in O} \left(t_{i+o}|_{\text{POS}} = Q_o \wedge t_{i+o}|_C = Y_o\right)$,
   „Bieżący segment ma część mowy P, kategoria gramatyczna C przyjmuje wartość X, a jego otoczenie ma ma konkretne wartości części mowy (Q_o) oraz jednej kategorii gramatycznej C (wartości Y_o)".

Wykorzystano również następujące szablony akcji:

1. $t_i := V$,
   „Zmień tag na V";
2. $t_i|_{\text{POS}} := R$,
   „Zmień część mowy na R, pozostawiając wartości zgodnych kategorii gramatycznych, a dla nowych wybierając dowolną wartość dopuszczoną przez analizator morfosyntaktyczny";
3. $t_i|_C := Z$,
   „Zmień wartość kategorii gramatycznej C na Z".

W obu typach szablonów:

- T, U, U_o, V – dowolny tag;
- U' – dowolny tag z poprzedniej fazy;
- P, Q, Q_o, R – dowolne klasy gramatyczne (części mowy);
- C – dowolna kategoria gramatyczna (uwzględniana w bieżącej fazie);
- X, Y, Y_o, Z – możliwe wartości kategorii gramatycznej C;
- $O \in \{\{1\}, \{-1\}, \{2\}, \{-2\}, \{1,2\}, \{-1,-2\}, \{1,2,3\}, \{-1,-2,-3\}\}$,
- zmienne O_1 oraz O_2, jak poprzednio, określają badane otoczenie bieżącego segmentu;
- zmienne P, Q, Q_o (dla wszystkich o jednocześnie), R, X, Y, Y_o (dla wszystkich o jednocześnie) i Z mogą mieć specjalną wartość $\star$, tzn. „dowolny".

Na koniec warto wspomnieć, że w drugiej fazie tagowania nie mają zastosowania te reguły, które spowodowałyby zmianę wyboru części mowy, przypadka lub osoby. Po prostu akcje nie są wykonywane, nawet jeśli predykat wskazuje, że reguła może być zastosowana w danym miejscu.

11.3.3. Leksykalne szablony transformacji

Kolejnym rozszerzeniem, które pojawiło się w literaturze (Brill 1994, Megyesi 1999), są szablony leksykalne, czyli takie, których predykaty umożliwiają identyfikację własności leksykalnych segmentów (u nas: końcówek i początków wyrazów). W PANTERZE zastosowano (poza wymienionymi powyżej) następujące szablony predykatów:

1a) $t_i = T \wedge$ i-ty segment kończy się na S';

1b) $t_i = T \wedge$ i-ty segment rozpoczyna się od S';

4a) $t_i|_{\text{POS}} = P \ \wedge \ t_i|_C = X \wedge$ i-ty segment kończy się na S
$\wedge \ \exists_{o \in O}\big(t_{i+o}|_{\text{POS}} = Q \wedge t_{i+o}|_C = Y\big)$,
„Bieżący segment kończy się na S, ma część mowy P, kategoria gramatyczna C przyjmuje wartość X, a jeden z sąsiednich (otoczenie O) segmentów ma część mowy Q, a wartością jego kategorii C jest Y";

4b) $t_i|_{\text{POS}} = P \wedge t_i|_C = X$
$\wedge \ \exists_{o \in O}\big(t_{i+o}|_{\text{POS}} = Q \wedge t_{i+o}|_C = Y$
$\wedge$ $(i+o)$-ty segment kończy się na $S\big)$,
„Bieżący segment ma część mowy P, kategoria gramatyczna C przyjmuje wartość X, a jeden z sąsiednich (otoczenie O) segmentów kończy się na S, ma część mowy Q, a wartością jego kategorii C jest Y";

gdzie S i S' to dowolne ciągi znaków, odpowiednio 3- i 2-literowe.

11.3.4. Optymalizacja wydajności

Kolejnym problemem do rozwiązania podczas prac nad NKJP była niska efektywność trenowania algorytmu Brilla, przynajmniej jeśliby zaimplementować go bezpośrednio tak, jak opisano w p. 11.2. Aby zwiększyć szybkość treningu, zaimplementowano zmodyfikowany algorytm FastTBL, zaproponowany w artykule Ngai i Floriana (2001). Jego istotą jest nieprzeliczanie ocen $good(r)$ i $bad(r)$ przy każdym przebiegu, lecz jedynie aktualizacja tych ocen, które mogły się zmienić w wyniku zmian w tagowaniu korpusu poprawek. Algorytm Ngai został przy okazji uproszczony, co opisano w pracy Acedańskiego (2010).

Wykonano także drugi krok – zrównoleglenie wykonywania zarówno treningu, jak i tagowania. Tager został przystosowany do pracy na maszynie wielowątkowej. Podczas trenowania zbiór szablonów transformacji jest dzielony między dostępne wątki wykonania. Każdy z nich utrzymuje wartość $good(r)$ i $bad(r)$ tylko dla transformacji powstałych z szablonów do niego przypisanych. Zrównoleglenie tagowania polega na podzieleniu zbioru plików wejściowych pomiędzy poszczególne wątki.

11.4. Ewaluacja

Ewaluacja tagera PANTERA została przeprowadzona na ręcznie anotowanym podkorpusie NKJP o rozmiarze ponad 1 200 000 segmentów. Do ewaluacji zastosowano metodę zaproponowaną w pracy Karwańskiej i Przepiórkowskiego (2011): korpus anotowany ręcznie podzielono na 10 części – 9 z nich stanowiło jednocześnie korpus treningowy oraz korpus poprawek, pozostała część była korpusem testowym. Przeprowadzono dziesięciokrotną walidację krzyżową, a podane wyniki są średnimi arytmetycznymi uzyskanych wartości.

Tabela 11.1. Wyniki ewaluacji – korpus NKJP (pełne tagi)

Czas (s)	**Liczba transformacji**		**Dokł.** (%)
	faza 1	faza 2	
1 567	1 638	602	92,93%

Miarą dokładności w tab. 11.1 jest procent segmentów, dla których tag całkowicie zgadza się z ręczną anotacją. Miara ta, jakkolwiek naturalna i szeroko stosowana, w taki sam sposób traktuje wszystkie rodzaje możliwych błędów tagowania – błędy w oznaczeniu części mowy są odzwierciedlone tak samo silnie, jak błędy w rozróżnieniu poszczególnych podrodzajów rodzaju męskiego. Dlatego też większa wartość miary dokładności nie musi zwiększać skuteczności

działania narzędzi wykorzystujących tager. Zagadnienie ewaluacji tagerów morfosyntaktycznych omawia szerzej praca Acedańskiego i Przepiórkowskiego (2010).

Czas działania programu podany w tab. 11.1 został uzyskany na maszynie wieloprocesorowej z procesorami AMD Opteron 2,3 GHz (0,5 MB pamięci cache). Wykorzystano sześć potoków obliczeniowych.

Aby przybliżyć charakterystykę błędów popełnianych przez PANTERĘ, dokładność tagowania w rozbiciu na poszczególne klasy i kategorie gramatyczne przedstawiono w tabelach 11.2–11.4. Nie zaskakuje to, że duże problemy sprawia rozróżnienie poszczególnych przypadków, szczególnie mianownika, biernika i dopełniacza. Widać także trudności z odróżnianiem poszczególnych rodzajów męskich zaimków i przymiotników (dla rzeczowników informacja o rodzaju jest podana przez analizator morfosyntaktyczny).

Warto również wspomnieć, że ponad 60% wszystkich wygenerowanych reguł zawiera predykaty leksykalne, a więc uwzględnia zazwyczaj końcówki tagowanych lub sąsiednich segmentów. Najczęściej reguły te służą do ujednoznaczniania przypadka lub do rozróżniania rzadko spotykanych klas gramatycznych, np. kublików i spójników.

Dane szczegółowe omówione powyżej i przedstawione w tabelach 11.2–11.5 pochodzą z pracy Acedańskiego (2010), opisującej ten sam tager, ale trenowany i ewaluowany na mniejszej części ręcznie anotowanego korpusu NKJP.

11.5. Instrukcja obsługi

Tager PANTERA jest opublikowany na licencji GPL. Strona projektu znajduje się pod adresem `http://code.google.com/p/pantera-tagger/`. Tam też można znaleźć pełną instrukcję obsługi. W szczególności znajduje się tam opis sposobu instalacji PANTERY. W tym miejscu przytaczamy najczęściej spotykane sytuacje użycia tagera.

11.5.1. Trenowanie tagera

Do trenowania PANTERY trzeba najpierw przygotować korpus treningowy. Do tego celu należy użyć anotowanego morfosyntaktycznie korpusu w formacie XCES. Trenowanie uruchamia się następującym poleceniem:

```
pantera --tagset nkjp --create-engine mój.btengine \
  --training-data dane_treningowe --threshold 6
```

W efekcie powstaje plik podany w wierszu poleceń (tutaj `mój.btengine`), w którym zapisane są transformacje oraz model tagera unigramowego. Parametr `threshold` określa minimalną jakość generowanych reguł (wartość $good(r) - bad(r)$).

Tabela 11.2. Błędy tagowania poszczególnych klas gramatycznych (pokazano wartości $> 0{,}01\%$)

Oczekiwana klasa	% segm.	Oczekiwana klasa	% segm.
subst	0,47%	prep	0,07%
qub	0,26%	pred	0,06%
adj	0,25%	num	0,05%
ger	0,21%	fin	0,04%
conj	0,10%	pact	0,03%
adv	0,09%	comp	0,03%
ppas	0,08%		

Tabela 11.3. Błędy tagowania poszczególnych kategorii gramatycznych (pokazano dla $> 0{,}01\%$)

Kategoria	% segm.
CASE	3,28%
GENDER	2,49%
NUMBER	0,72%
ASPECT	0,06%
ACCOMMODABILITY	0,03%

Tabela 11.4. Błędy dla konkretnych wartości kategorii gramatycznych

Oczekiwane	Przypisane przez tager	% segm.
CASE(NOM)	CASE(ACC)	1,11%
CASE(ACC)	CASE(NOM)	0,73%
GENDER(M1)	GENDER(M3)	0,39%
CASE(GEN)	CASE(ACC)	0,39%
NUMBER(SG)	NUMBER(PL)	0,38%
NUMBER(PL)	NUMBER(SG)	0,34%
GENDER(M3)	GENDER(M1)	0,31%
GENDER(M3)	GENDER(N)	0,26%
GENDER(M1)	GENDER(F)	0,21%
GENDER(F)	GENDER(M3)	0,19%
GENDER(M3)	GENDER(F)	0,19%
GENDER(F)	GENDER(N)	0,17%
CASE(GEN)	CASE(NOM)	0,17%
CASE(ACC)	CASE(GEN)	0,15%
GENDER(N)	GENDER(M3)	0,14%

Tabela 11.5. Przykłady reguł leksykalnych wygenerowanych przez PANTERĘ w pierwszej fazie

Nr	r	$good(r)$	$bad(r)$
3	Zmień wartość przypadka wymaganego przez przyimek z biernika na miejscownik, jeśli wyraz kończy się na `na` (w praktyce reguła dotyczy po prostu przyimka *na*) oraz jeśli jeden z dwóch kolejnych segmentów jest w bierniku.	2 496	113
7	Zmień przypadek przymiotnika z miejscownika na narzędnik, jeśli jeden z trzech kolejnych segmentów jest w narzędniku oraz kończy się na `em`.	921	29

Jako dane treningowe można przekazać ścieżkę do katalogu bądź też pojedyncze pliki, które zostaną połączone w jeden zbiór danych treningowych.

11.5.2. Uruchomienie tagera

Do uruchomienia tagera potrzebne są transformacje wyprodukowane podczas trenowania, jak również statystyki wykorzystywane przez tager unigramowy. Informacje te zapisywane są w pliku o rozszerzeniu `.btengine` podczas trenowania. W repozytorium kodu PANTERY, w podfolderze `engines/` można znaleźć taki plik wytrenowany na ręcznie anotowanej części korpusu NKJP. Oczywiście można również samemu przygotować taki plik, według opisu powyżej.

Dane wejściowe do tagowania mogą być w różnych formatach:

1. w formacie TEI NKJP – pliki wejściowe muszą się wówczas nazywać `text_structure.xml`;
2. w formacie XCES IPI PAN – pliki z rozszerzeniem `.xml` z odpowiednim nagłówkiem,
3. pliki tekstowe, z rozszerzeniem `.txt`;

Teraz należy uruchomić tager w następujący sposób:

```
pantera --tagset nkjp --engine mój.btengine pliki_foldery ...
```

Można przekazać jako argumenty pojedyncze pliki wejściowe bądź nazwy folderów (wówczas przeszukiwane są wszystkie podfoldery w celu znalezienia plików rozpoznawanych jako wejściowe).

Domyślnie tager przeprowadzi podział na zdania, analizę morfoskładniową, dezambiguację segmenacji oraz tagowanie. Wykonywanymi fazami można sterować przy użyciu opcji wiersza poleceń (patrz `pantera --help`).

11.5.3. Uruchomienie trybu ewaluacji

Jeżeli podczas uruchomienia tagera poda się mu na wejście pliki zawierające już ujednoznaczniony opis morfoskładniowy (co jest w bieżącej wersji możliwe jedynie dla formatu XCES IPI PAN przy użyciu XML-owych atrybutów `disamb`), oprócz ponownego otagowania podanych tekstów, przeprowadzona zostanie ewaluacja. Wyniki ogólne zostaną wypisane na wyjście, a szczegółowe zostaną zapisane do pliku z rozszerzeniem `.errors.txt` i nazwie odpowiadającej nazwie pliku wejściowego.

11.5.4. Inne opcje

Oprócz funkcji wspomnianych powyżej, warto wiedzieć o kilku dodatkowych możliwościach PANTERY. Informacje o ich sposobie użycia można znaleźć, uruchamiając tager bez podania argumentów.

1. Istnieje możliwość definiowania własnych słowników morfosyntaktycznych. Interpretacje tam znalezione są używane zamiast tych zwróconych przez analizator Morfeusz.
2. Obsługiwany jest tryb tworzenia skompresowanych plików wyjściowych.
3. Jest opcja tagowania jedynie tych plików, dla których nie istnieje plik wynikowy. Ułatwia to kontynuację przetwarzania dużego zbioru danych w przypadku jego przerwania.

11.6. Wnioski końcowe

Tager PANTERA użyty w projekcie NKJP jest już drugim obok TaKIPI tagerem dla języka polskiego dostępnym na otwartej licencji. Oba rozwiązania wykorzystują bardzo różne algorytmy dezambiguacji i uzyskują podobną jakość tagowania. W przeciwieństwie do PANTERY, TaKIPI oprócz statystyki wykorzystuje informacje lingwistyczne podane w formie ręcznie pisanych reguł. Można więc podejrzewać, że połączenie tych tagerów mogłoby zauważalnie zwiększyć jakość tagowania.

Jeśli zaś chodzi o samą PANTERĘ, to nie wykorzystano jeszcze wszystkich możliwości jej rozwoju. W szczególności praca Acedańskiego (2010) podaje przykłady zdań, w których nie można wykonać dezambiguacji niektórych słów, patrząc jedynie na trójsegmentowy kontekst. Praca nad rozszerzeniem kontekstu szablonów mogłaby więc przynieść pozytywne rezultaty.

Mimo to już teraz PANTERA jest systemem w pełni funkcjonalnym i gotowym do wykorzystania nie tylko w projekcie NKJP.

12

Automatyczne znakowanie sensami słów

Mateusz Kopeć, Rafał Młodzki, Adam Przepiórkowski

12.1. Wprowadzenie

12.1.1. Rozróżnianie sensów słów

Trywialne jest stwierdzenie, że w języku naturalnym pojedyncze słowo może mieć wiele znaczeń. Na przykład w języku polskim słowo „zamek" oznacza m.in. budowlę z kamienia (sens 1) lub urządzenie umożliwiające zamykanie czegoś (sens 2). System do *rozróżniania sensów słów* (ang. *Word Sense Disambiguation*; WSD) automatycznie decyduje, które ze znaczeń homonimicznego słowa wystąpiło w danym kontekście, np.:

(12.1) *Zamek*$_2$ w drzwiach trzeba było wymienić.

(12.2) Okazały *zamek*$_1$ stał na wzgórzu.

W pierwszym zdaniu opisany jest mechanizm w drzwiach, w drugim natomiast budowla. Te dwa sensy zostały oznaczone za pomocą liczb w indeksie dolnym.

Dla człowieka takie rozróżnienie jest zadaniem nietrudnym, na ogół wykonywanym zupełnie nieświadomie, systemowi komputerowemu może jednak nastręczyć wiele trudności.

Zadanie systemu rozróżniającego sensy słów polega na automatycznym oznaczaniu występujących w kontekście słów wieloznacznych właściwymi numerami sensów. W tym celu potrzebna jest baza tekstów do oznaczenia oraz odpowiedni słownik, zawierający słowa mające więcej niż jedno znaczenie, a dodatkowo dostarczający pulę sensów do wyboru dla każdego takiego słowa.

12.1.2. Motywacja

Skuteczny system WSD jest potencjalnie przydatny w wielu dziedzinach przetwarzania języka naturalnego, w tym w:

1. tłumaczeniu maszynowym,
2. ekstrakcji informacji,
3. automatycznym odpowiadaniu na pytania itp.

Prosty przykład tłumaczenia maszynowego może wyglądać następująco: wyobraźmy sobie, że w tłumaczonym z języka polskiego na język angielski zdaniu występuje słowo „piłka". Bez podania jego sensu system może błędnie przetłumaczyć to słowo, ponieważ może ono oznaczać zarówno małą piłę do cięcia, jak i okrągły przedmiot pożądania 22 mężczyzn biegających po boisku. W wielu językach obcych (np. w angielskim) obu tym sensom odpowiadają jednak różne tłumaczenia, stąd konieczność ujednoznacznienia tłumaczonego słowa[1].

Warto też uzasadnić potrzebę zbadania zagadnienia WSD w kontekście języka polskiego. Są ku temu dwa zasadnicze powody. Po pierwsze, zdecydowana większość prac z dziedziny WSD odwołuje się niemal wyłącznie do języka angielskiego. Po drugie, nie istnieje publicznie dostępny, duży korpus z tekstami polskojęzycznymi, zawierający anotację sensami słów. Brak również narzędzi umożliwiających badanie tego zagadnienia dla języka polskiego.

12.1.3. Narodowy Korpus Języka Polskiego

Ze względu na wymienione powody uzasadnione było uwzględnienie w projekcie Narodowego Korpusu Języka Polskiego badania rozróżniania sensów słów. Wśród wielu poziomów anotacji Korpusu znalazła się również anotacja sensami słów, zarówno stworzona ręcznie przez zespół anotatorów (dla podkorpusu milionowego; zob. rozdz. 5), jak i wygenerowana przez automatyczne narzędzia (dla pełnego NKJP).

12.2. Opis projektu

12.2.1. Słownik sensów

Jednym z kluczowych aspektów przy definiowaniu zadania WSD jest wybór słownika zawierającego wyliczenia sensów poszczególnych słów. Nie każdy słownik nadaje się do tego celu – najistotniejszą i najbardziej pożądaną cechą jest

[1] W praktyce współczesne statystyczne systemy tłumaczenia maszynowego zwykle same dokonują ujednoznacznienia słów, więc nie potrzebują korzystać z takiego wyspecjalizowanego modułu zewnętrznego.

„gruboziarnistość”[2] sensów w nim zdefiniowanych. Jest oczywiste, że sytem komputerowy prawdopodobnie nie podoła przypisania słowu „mieć” jednego z 14 znaczeń podanych w standardowym słowniku. Ponadto trudno sobie wyobrazić praktyczne zastosowanie wyników działania tak precyzyjnego dezambiguatora. W środowisku WSD przyjęło się za standard korzystanie z „gruboziarnistych” słowników, najczęściej dostarczających 2–3 wyraźnie różne znaczenia danego słowa (Agirre i Edmonds 2006).

Ponieważ nie udało się znaleźć odpowiedniego do tego zadania słownika, został on stworzony specjalnie na potrzeby tego projektu (zob. rozdz. 7). W związku z badawczą naturą projektu zakres utworzonego zasobu jest ograniczony, słownik obejmuje jedynie 106 wieloznacznych, często występujących w języku polskim słów. Słownik sensów stworzono w oparciu o *Multimedialny słownik szkolny PWN* (Bańko 2005; jest to elektroniczna wersja *Innego słownika języka polskiego*; Bańko 2000). Zredukowaną liczbę sensów otrzymano przez łączenie bardziej szczegółowych znaczeń w grupy znaczeniowo podobne. W powstałym słowniku jedno słowo ma średnio 2,85 sensów.

12.2.2. Zakres projektu i przykłady treningowe

Główny cel projektu to oznaczenie całego Narodowego Korpusu Języka Polskiego sensami słów, czyli odnalezienie wszystkich wystąpień każdego spośród 106 słów zdefiniowanych w nowoutworzonym słowniku, a następnie przypisanie sensu odpowiedniego w danym kontekście. Ze względu na rozmiar korpusu niemożliwością było dokonanie takiej anotacji ręcznie, koniecznością było stworzenie automatycznego dezambiguatora, z możliwie największą skutecznością dokonującego prawidłowych wyborów.

Obecnie najlepsze systemy WSD działają w oparciu o algorytmy uczenia maszynowego, dlatego oczywistym wyborem było użycie tego podejścia w projekcie. Uczenie maszynowe tzw. z nadzorem (ang. *supervised machine learning*) wykorzystuje możliwie największe ilości ręcznie anotowanych przykładów w celu wyuczenia klasyfikatora, potrafiącego (dzięki wiedzy zdobytej podczas procesu uczenia) zaklasyfikować nowy, nieanotowany przykład. Aby wykorzystać tę technikę, konieczne było stworzenie zbioru przykładów treningowych, co zostało wykonane przez ręczne zaanotowanie podkorpusu NKJP zawierającego około miliona segmentów, w tym ponad 34 tysiące wystąpień słów wieloznacznych.

Anotowanie korpusu treningowego było dokonywane przez dwóch niezależnych anotatorów (por. p. 6.6.2). Na 34 205 wystąpienia słów wieloznacznych, ich

[2] „Gruboziarnistość” podziału słowa na sensy oznacza ich niewielką liczbę, w przeciwieństwie do „drobnoziarnistego” podziału, który wyszczególnia dużą liczbę sensów.

oznaczenia sensów były zgodne w 32 321 przypadków, co daje współczynnik 95% zgodności. Jest to górna granica skuteczności systemu WSD dokonującego tego samego procesu. Tę 5-procentową różnicę w oznaczeniach stanowią przypadki kontrowersyjne, które zostały uzgodnione przed stworzeniem ręcznie anotowanego korpusu.

Stworzenie podkorpusu ręcznie anotowanego sensami było użyteczne także z dwóch innych powodów. Po pierwsze, taki zasób jest konieczny, aby ocenić skuteczność systemu WSD. Pozwala porównać anotację człowieka (traktowaną jako wzorcowa) z anotacją wygenerowaną przez program. Po drugie, dzięki takiemu zasobowi możliwe jest przetestowanie innych metod rozróżniania sensów słów nie uwzględnionych w tym projekcie.

12.2.3. Struktura projektu

O ile wybór podejścia z nadzorem do tworzenia systemu WSD nie budził wątpliwości, o tyle wybór konkretnej metody stanowił problem do rozwiązania. W tym celu zostało użyte narzędzie *Word Sense Disambiguation Development Environment* (Młodzki i Przepiórkowski 2009), dalej określane jako WSDDE, środowisko pozwalające przeprowadzać eksperymenty WSD przy użyciu różnych rodzajów metod uczenia maszynowego. Za jego pomocą uzyskane zostały eksperymentalnie wybrane najskuteczniejsze metody automatycznej dezambiguacji sensów słów, użyte do otagowania całego Narodowego Korpusu Języka Polskiego.

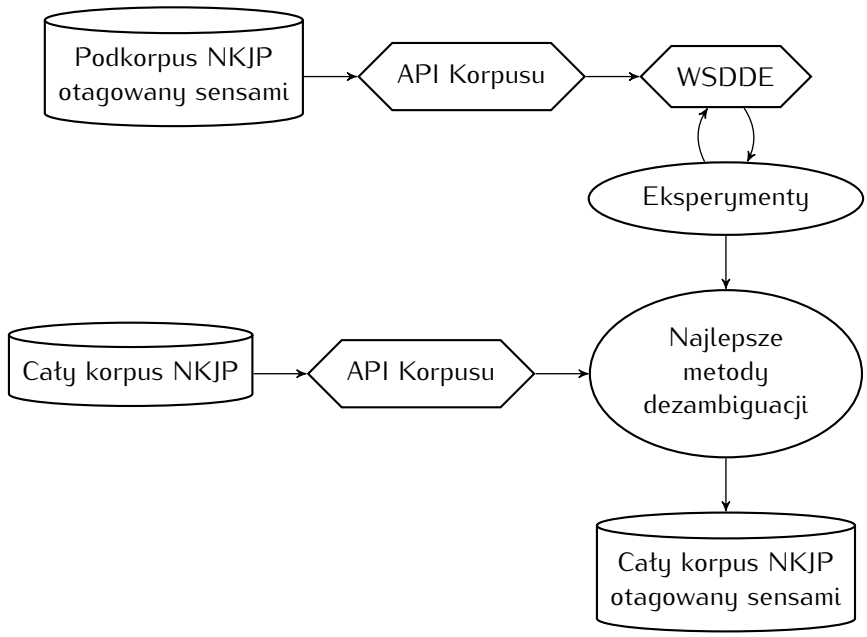
Zanim użyto WSDDE, należało utworzyć API[3] dla korpusu, ponieważ bezpośrednie przetwarzanie danych przechowywanych w ogromnej ilości plików XML nie było wygodne z punktu widzenia programisty. Dodatkową zaletą tego podejścia jest udostępnienie API korpusu do innych zastosowań.

Po stworzeniu wygodnego dostępu do danych, wykorzystane mogło zostać nieco zmodyfikowane środowisko Word Sense Disambiguation Development Environment. Za jego pomocą przeprowadzono tysiące eksperymentów, pozwalających na wybranie najlepszej metody dezambiguującej dla każdego z rozpatrywanych słów. Przy użyciu tych metod wszystkie wystąpienia 106 wieloznacznych słów w korpusie NKJP zostały automatycznie otagowane najbardziej prawdopodobnymi sensami.

Największym wyzwaniem było osiągnięcie możliwie najlepszej skuteczności dezambiguacji, przy zachowaniu najmniejszej złożoności obliczeniowej, ze względu na rozmiar danych do przetworzenia (głównie w ostatniej fazie, przy tagowaniu całości).

[3] Application Programming Interface – interfejs programowania aplikacji. W tym wypadku jest to warstwa pośrednia, udostępniająca precyzyjnie określone metody dostępu do danych.

Rysunek 12.1. Struktura projektu



12.3. Word Sense Disambiguation Development Environment

Word Sense Disambiguation Development Environment to środowisko umożliwiające przeprowadzanie eksperymentów WSD i mierzenie skuteczności różnych metod z nadzorem. Pozwala ono na określenie korpusu testowego i treningowego oraz wybór cech kontekstu, które mają być brane pod uwagę. Ponadto można zdefiniować algorytmy automatycznej selekcji cech oraz algorytm uczenia maszynowego, które to powinny być zgodne interfejsami w narzędzia WEKA (zob. Witten i Frank 2005), udostępniającym liczne algorytmy maszynowego uczenia się.

Wszystkie możliwe kombinacje powinny zostać przetestowane w celu znalezienia najlepszych metod. Korpusem tesowym i treningowym był podkorpus NKJP anotowany ręcznie.

Ekstrakcja cech z kontekstu dokonywana była za pomocą czterech rodzajów generatorów cech, wydobywających:

1. cechy tematyczne (ang. *thematic features*, TF),
2. cechy strukturalne I (ang. *structural features* I, SF1),
3. cechy strukturalne II (ang. *structural features* II, SF2),
4. cechy słowa kluczowego (ang. *keyword features*, KF).

Każdą z cech opisano poniżej.

12.3.1. Cechy tematyczne (TF)

Cechy tematyczne mają za zadanie dostarczać informacji na temat ogólnego znaczenia kontekstu. Mówią one o obecności konkretnych słów w określonym przedziale zawierającym rozpatrywane słowo wieloznaczne. Ekstrakcją tych cech z kontekstu steruje się za pomocą następujących parametrów:

1. Wielkość okna – liczona w prawo i w lewo od dezambiguowanego słowa. Zatem wartość 20 oznacza, że będzie brane pod uwagę 40 słów, jeżeli dostępny kontekst ma wystarczającą szerokość.
2. Lematyzacja (sprowadznie do formy podstawowej, nieodmienionej) słów z kontekstu – 0 oznacza wyłączoną, 1 włączoną.
3. Binarność – 1 oznacza, że ma być brana pod uwagę tylko obecność danego słowa w kontekście lub jej brak, 0 oznacza zliczanie wystąpień.

Poniższy przykład pokazuje, jak wyglądałyby cechy tematyczne dla słowa *zamek* (zdania (12.1)–(12.2) w p. 12.1.1). Wiersze oznaczają wektory cech wygenerowane dla tych dwóch zdań, kolumny zaś oznaczają poszczególne cechy kontekstów.

(12.3)

Kontekst	drzwi	trzeba	stać	wzgórze	…
(12.1)	1	1	0	0	…
(12.2)	0	0	1	1	…

W tym przykładzie użyto lematyzacji, zatem przed wydobyciem cech kontekstu najpierw sprowadzono słowa do form podstawowych. Rozpatrywana jest jedynie obecność słowa w kontekście, zatem liczba 1 oznacza, że słowo się pojawiło, liczba 0 zaś mówi o jego braku. Przykładowo, w pierwszej kolumnie, oznaczonej słowem *drzwi*, zawarta jest informacja, że to słowo znalazło się w pierwszym kontekście, natomiast w drugim nie.

12.3.2. Cechy strukturalne I (SF1)

Cechy strukturalne typu pierwszego dostarczają informacji o obecności słowa w kontekście, podobnie jak cechy tematyczne. Biorą jednak dodatkowo pod uwagę to, w jakiej odległości względem słowa kluczowego występuje dane słowo, a także czy stoi przed nim, czy za nim.

Przetestowane parametry cech strukturalnych są podobne jak poprzednio:

1. Wielkość okna (do 5) – liczona w prawo i w lewo od dezambiguowanego słowa.
2. Lematyzacja – 0 oznacza wyłączoną, 1 włączoną.
3. Binarność – 1 oznacza, że ma być pod uwagę brana tylko obecność danego słowa w kontekście lub jej brak, 0 oznacza ciągłą wartość (miarę IDF).

Rozpatrywany rozmiar przedziału po obu stronach słowa kluczowego jest mniejszy (do 5 segmentów), ponieważ jest mało prawdopodobne, aby istniała istotna ściśle pozycyjna zależność o tak dużym zasięgu. Poniższy przykład prezentuje wektory cech strukturalnych I:

(12.4)

Kontekst	drzwi+2	stał+1	w+1	Okazały-1	…
(12.1)	1	0	1	0	…
(12.2)	0	1	0	1	…

Znaczenie nazw cech należy interpretować w następujący sposób: *drzwi+2* oznacza obecność słowa *drzwi* dwa segmenty za słowem kluczowym, natomiast *Okazały-1* mówi o poprzedzającym słowo kluczowe wyrazie *Okazały*. W tym przykładzie nie użyto lematyzacji (zob. *stał+1*, a nie *stać+1*).

12.3.3. Cechy strukturalne II (SF2)

Cechy strukturalne II różnią się od poprzednich cech tym, że zamiast wydobywać obecność słów na danej pozycji w kontekście, badają wyłącznie ich interpretację morfosyntaktyczną.

Oto ich parametry:

1. Wielkość okna – liczona w prawo i w lewo od dezambiguowanego słowa.
2. Binarność – 1 oznacza, że ma być pod uwagę brana tylko obecność danego słowa w kontekście lub jej brak, 0 oznacza ciągłą wartość (IDF).
3. Proste klasy gramatyczne – 1 oznacza wybór prostych klas gramatycznych, 0 wybór drobniejszego podziału klas gramatycznych (więcej informacji w dodatku 12.A).
4. Rozszerzona morfoskładnia – 1 sygnalizuje obecność innych informacji o interpretacji morfosyntaktycznej (np. liczba, rodzaj, przypadek), 0 oznacza ich brak.

(12.5)

Kontekst	noun+2	verb+1	prep+1	adj-1	…
(12.1)	1	0	1	0	…
(12.2)	0	1	0	1	…

12.3.4. Cechy słowa kluczowego (KF)

Ostatnim rodzajem wydobywanych cech były cechy samego słowa wieloznacznego, czyli określenie jego interpretacji morfoskładniowej, a także ustalenie, czy rozpoczyna się ono wielką literą.

Dostępne parametry generowania cech słowa kluczowego:

1. Proste klasy gramatyczne – 1 oznacza wybór prostych klas gramatycznych, 0 wybór drobniejszego podziału klas gramatycznych (więcej informacji w dodatku p. 12.A).
2. Rozszerzona morfoskładnia – 1 sygnalizuje obecność innych informacji o interpretacji morfosyntaktycznej (np. liczba, rodzaj, przypadek), 0 oznacza ich brak.
3. Wielka litera – 1 oznacza użycie cechy sprawdzającej obecność wielkiej litery na początku słowa, 0 jej pominięcie.

(12.6)

Kontekst	noun	verb	singular	capitalized	…
(12.1)	1	0	1	1	…
(12.2)	1	0	1	0	…

12.3.5. Selekcja cech i wybór algorytmu uczenia maszynowego

Z kontekstu można wydobyć ogromne ilości tych czterech typów cech. Aby umożliwić algorymowi uczenia maszynowego skuteczną naukę, należy dokonać selekcji cech, pozostawiając wyłącznie te, które niosą pewną informację na temat sensu dezambiguowanego słowa. WSDDE umożliwia użycie automatycznych algorytmów selekcji cech. Przetestowane zostały różne konfiguracje dwóch dostępnych w WEKA algorytmów: InfoGain (badającego przyrost informacji związany z daną cechą) i CfsSubsetEval (analizującego redundancję w podzbiorach cech).

Po odsianiu nieistotnych cech kontekstu algorytm uczenia maszynowego może zająć się tworzeniem klasyfikatora, rozróżniającego sensy danego słowa. Sprawdzono różne algorytmy, dostępne w bibliotece WEKA:

– NaiveBayes (naiwny klasyfikator Bayesowski),
– J48graft (drzewo decyzyjne C4.5),
– VFI (Voting Feature Intervals),
– KStar (wariant metody k-najbliższych sąsiadów),
– BayesNet (sieć Bayesowska),
– DecisionTable (tablica decyzyjna),
– RandomForest (las losowy),
– AdaBoostM1 (wzmacnianie adaptacyjne).

12.4. Przeprowadzone eksperymenty

12.4.1. Korpus treningowy i testowy

Za korpus treningowy posłużyła wstępna wersja ręcznie anotowanego podkorpusu NKJP, zawierającego 3889 tekstów, w których znalazło się 1 217 822 segmentów, a wśród nich 34 186 wystąpień słów wieloznacznych. Szczegółowe statystyki dotyczące wystąpień poszczególnych słów wieloznacznych znajdują się w dodatku 12.B na końcu rozdziału.

12.4.2. Ewaluacja

W celu jak najlepszego wykorzystania ograniczonego ręcznie anotowanego zbioru przykładów wykorzystana została walidacja krzyżowa, w dwóch odmianach: *10-fold* i *leave-one-out*. Przy walidacji *10-fold* zastosowano podział losowy, ale zawsze taki sam dla każdej testowanej metody.

Dla porównania, prosta heurystyka wybierania zawsze najczęstszego sensu dla tego korpusu dawała wynik 78,3%.

12.4.3. Eksperymenty

W tab. 12.1 znajduje się opis przeprowadzonych eksperymentów. Każda z metod danego eksperymentu była testowana osobno dla każdego spośród 106 słów wieloznacznych. W każdym z tych przypadków wybierana była najskuteczniejsza metoda, czyli taka, która zastosowana do danego słowa wieloznacznego osiągnęła największą liczbę poprawnych odpowiedzi.

Wynik całego eksperymentu stanowi procent wystąpień leksemów wieloznacznych, które zostały dobrze ujednoznacznione, przy założeniu, że dla każdego słowa wieloznacznego używana jest najskuteczniejsza dla niego metoda spośród testowanych w danym eksperymencie.

Łącznie przetestowano 215 metod dla każdego ze słów wieloznacznych, co daje razem 22 790 sprawdzonych metod.

12.4.4. Statystyki dotyczące najlepszych metod

Poniżej zaprezentowano metody, które znajdują się wśród ostatecznie wybranych w eksperymencie 5B najlepszych algorytmów dla każdego spośród 106 słów wieloznacznych.

Tabela 12.1. Opis przeprowadzonych eksperymentów

Nr	Liczba metod	Opis	Wynik
1	8	Naive Bayes, filtrowanie InfoGain+CFSSubsetEval. Zmienna wielkość okna TF: 10, 15, 20, 25 i zmienna obecność pakietu cech: SF2(2,1,1,0), SF1(2,1,1), KF(1,1,1)	0,8974
2	16	jak w eksperymencie 1, lecz z oknem TF o mniejszych rozmiarach: 0, 1, 2, 5, 6, 7, 8, 9	0,8948
3	4	Naive Bayes, filtrowanie InfoGain+CFSSubsetEval, ze zmienną wielkością okna TF: 10, 15, 20, 25 i stałą obecnością pakietu cech: SF2(2,1,1,1), SF1(2,1,1), KF(1,1,1)	0,8986
4	72	Naive Bayes, filtrowanie InfoGain+CFSSubsetEval, stały TF(10,1,1), zmienne SF1(1/2/3,1,1/0), zmienne SF2(1/2/3,1,1/0,1), zmienne KF(1/0,1,1).	0,8988
5	7	filtrowanie InfoGain+CFSSubsetEval, 7 różnych algorytmów: J48graft, VFI, KStar, BayesNet, DecisionTable, RandomForest, AdaBoostM1; cechy stałe: TF(10,1,1), SF1(1,1,1), SF2(1,1,1,1), KF(1,1,1)	0,8963
6	108	4 rodzaje filtrowania, 3 algorytmy: J48, KStar, BayesNet; generatory: stałe TFG(10,1,1), zmienne SF1(1/2/3,1,1), zmienne SF2(1/2/3,1,1,1), stałe KF(1,1,1)	0,9121
W	215	wszystkie powyższe metody	0,9148
5B	5	5 najlepszych dla każdego leksemu z ewaluacją *leave-one-out*	0,9146

Algorytm uczenia maszynowego

Jak pokazuje tab. 12.2, najbardziej popularnym algorytmem jest naiwny klasyfikator Bayesowski, który został wykorzystany w około 50% przypadków.

W wypadku trzech słów deterministyczny klasyfikator osiąga 100% skuteczności. Dzieje się tak dlatego, że wśród przykładów treningowych występuje tylko jeden sens. *Trywialny* jest takim właśnie klasyfikatorem.

Selekcja cech

Statystyki użycia algorytmów selekcji cech zostały przedstawione w tab. 12.3 (trzy przypadki z klasyfikatorem *trywialnym* pominięto):

Najpopularniejszą metodą selekcji atrybutów okazało się filtrowanie najbardziej istotnych cech za pomocą badania przyrostu informacji (InfoGain), a następnie rozpatrzenie różnych ich podzbiorów (CfsSubsetEval). Z przykładów treningowych wygenerowanych zostało około 2565 cech na słowo, a po etapie

Tabela 12.2. Wykorzystanie algorytmów uczenia maszynowego

Algorytm uczenia	Liczba metod
NaiveBayes	51
BayesNet	32
KStar	10
J48graft	7
DecisionTable	2
RandomForest	1
Trywialny	3
Σ	106

Tabela 12.3. Wykorzystanie algorytmów selekcji cech

Algorytm selekcji	Liczba użyć
InfoGain+CfsSubsetEval	60
InfoGain	43
Σ	103

filtrowania cech pozostało ich średnio 117. Tabela tab. 12.4 zawiera listę najczęstszych cech kontekstu, które są istotne przy dezambiguacji sensu słowa.

Ogólne statystyki wykorzystania po filtrowaniu cech poszczególnych typów zaprezentowane są w tab. 12.5.

Cechy tematyczne stanowią zdecydowaną większość branych pod uwagę cech. Nie świadczy to jednak o ich przewadze co do niesionej wartości informacyjnej, ponieważ liczba cech tego typu przed filtrowaniem również była największa.

Przykład klasyfikatora

Przykład klasyfikatora pokazano w oparciu o wieloznaczne słowo *uwaga*, które może być rozumiane w dwojaki sposób:

1. skupienie myśli lub zainteresowań na czymś,
2. spostrzeżenia lub komentarze, krytyka.

Algorytm selekcji cech zredukował ich liczbę z 2590 do 17. W tab. 12.6 przedstawiono pięć spośród nich, niosących największą informację o sensie słowa.

12.4.5. Podsumowanie

Opisany w niniejszym rozdziale projekt oznaczenia sensów słów w NKJP jest bezprecendesowy dla języka polskiego pod względem liczby tekstów użytych do

Tabela 12.4. Najczęściej wykorzystywane cechy

Liczba użyć	Cecha	Typ cechy
31	m3+1	SF2
31	w	TF
31	interp+1	SF2
31	capitalized	KF
32	sg+1	SF2
33	pl+1	SF2
33	prep+1	SF2
33	gen	KF
34	się	TF
38	gen+1	SF2
40	noun+1	SF2
40	sg	KF
40	pl	KF
40	prep-1	SF2

Tabela 12.5. Liczby cech poszczególnych typów po etapie selekcji

Typ cechy	Liczba użyć
TF	7 436
SF1	2 026
SF2	2 163
KF	795

uczenia i ewaluacji. Zapewne będzie on więc stanowił istotny punkt odniesienia podczas przyszłych badań tego zagadnienia, dodatkowo może posłużyć również jako źródło korpusu do ewaluacji przyszłych metod.

Wyniki uzyskane przez najlepsze spośród testowanych algorytmów potwierdzają wysoką skuteczność metod opartych na algorytmach uczenia maszynowego do rozwiązywania problemu automatycznej dezambiguacji sensów słów. Metody z nadzorem oczywiście mają jednak również swoje ograniczenia; jednym z nich jest wysoki nakład pracy konieczny do stworzenia zasobów treningowych do uczenia. Z tego powodu przyszłe badania automatycznego znakowania sensami słów z pewnością będą obejmować również metody niekorzystające z ręcznie utworzonych przykładów treningowych. Jednakże nawet takie projekty mogą skorzystać z zasobów NKJP w celu sprawdzenia skuteczności rozwiązań oraz porównania ich do przedstawionych w tym rozdziale wyników.

Tabela 12.6. Najlepsze cechy dla słowa *uwaga*

Cecha	Typ cechy
pl	KF
o+1	SF1
zgłaszać	TF
do+1	SF1
skoro	TF

12.A. Dodatek. Drobne i proste klasy gramatyczne

Przez „drobne klasy gramatyczne" rozumiemy tutaj klasy gramatyczne opisane w rozdz. 6. „Proste klasy" są wynikiem agregacji dwóch podzbiorów „drobnych klas" w nowe klasy „noun" i „verb":

1. verb = {pact, ppas, winien, praet, bedzie, fin, impt, aglt, ger, imps, inf, pant, pcon},
2. noun = {subst, depr, xxs, ger, ppron12, ppron3}.

12.B. Dodatek. Dane ręcznie anotowanego korpusu

W tab. 12.7 zaprezentowane są statystyki wystąpień słów wieloznacznych w ręcznie anotowanym podkorpusie NKJP. Liczba sensów może odbiegać od liczby sensów w użytym słowniku, ponieważ oznacza sensy, które wystąpiły w korpusie. Może zatem być mniejsza, jeżeli nie wszystkie sensy danego słowa są reprezentowane. Może być też większa, ponieważ anotatorzy mieli możliwość dokonania anotacji sensem *INNY* w przypadkach, gdy żadna z definicji dostępnych w słowniku sensów nie odpowiadała użyciu słowa.

Kolumna *MFS* odpowiada wynikowi heurystyki najczęstszego sensu dla danego słowa.

Tabela 12.7. Statystyka ręcznie oznaczonych wystąpień słów wieloznacznych

Słowo	Liczba wystąpień	Liczba sensów	Dystrybucja sensów	MFS
akcja	369	2	257/112	0,696
bliski	279	2	196/83	0,703
brać	299	7	142/98/30/17/6/4/2	0,475
chodzić	675	4	401/270/3/1	0,594
ciało	256	3	238/17/1	0,93
członek	368	3	363/3/2	0,986
czuć	354	2	352/2	0,994

Słowo	Liczba wystąpień	Liczba sensów	Dystrybucja sensów	MFS
dać	578	4	364/211/2/1	0,63
dawać	273	3	237/35/1	0,868
dodać	168	2	164/4	0,976
dokonać	208	2	197/11	0,947
dostać	338	4	267/40/18/13	0,79
doświadczenie	125	2	111/14	0,888
drogi	130	2	71/59	0,546
działanie	265	2	261/4	0,985
film	336	2	332/4	0,988
forma	255	3	239/9/7	0,937
góra	274	5	125/104/26/14/5	0,456
grać	240	4	104/70/55/11	0,433
izba	177	2	101/76	0,571
język	258	2	236/22	0,915
kierować	195	4	123/47/18/7	0,631
klasa	197	4	105/58/29/5	0,533
kolej	125	2	97/28	0,776
kultura	276	2	275/1	0,996
letni	53	3	43/7/3	0,811
leżeć	262	3	258/3/1	0,985
liczyć	264	2	160/104	0,606
materiał	196	3	108/67/21	0,551
mina	41	2	35/6	0,854
nastąpić	137	1	137	1,0
objąć	134	3	73/49/12	0,545
oddział	197	2	138/59	0,701
okazać	300	2	294/6	0,98
okres	375	2	367/8	0,979
państwo	715	2	569/146	0,796
piłka	78	1	78	1,0
pismo	173	2	142/31	0,821
podać	255	3	251/3/1	0,984
podjąć	255	2	248/7	0,973
podnieść	166	3	142/20/4	0,855
podstawa	279	2	277/2	0,993
pokój	253	2	193/60	0,763
pole	120	4	88/22/9/1	0,733
powód	322	2	317/5	0,984
powstać	235	2	232/3	0,987
pozostać	159	3	101/45/13	0,635
pozostały	152	3	128/21/3	0,842

Słowo	Liczba wystąpień	Liczba sensów	Dystrybucja sensów	MFS
pozycja	105	3	71/18/16	0,676
program	520	3	318/109/93	0,612
prosty	516	2	499/17	0,967
prowadzenie	7	1	7	1,0
prowadzić	632	4	493/112/21/6	0,78
przeprowadzić	215	4	200/12/2/1	0,93
przyznać	220	2	156/64	0,709
punkt	323	4	101/97/71/54	0,313
rada	599	3	528/37/34	0,881
raz	1 422	3	1 357/63/2	0,954
rola	250	2	248/2	0,992
rynek	267	2	205/62	0,768
rząd	576	2	521/55	0,905
sam	1 590	4	1 220/264/97/9	0,767
skład	137	3	122/13/2	0,891
składać	175	6	78/68/12/11/4/2	0,446
sprawa	1 502	3	1 337/164/1	0,89
stać	849	5	417/387/38/6/1	0,491
stan	523	4	446/42/31/4	0,853
stanąć	175	3	75/62/38	0,429
stanowisko	315	2	194/121	0,616
stopień	207	3	166/31/10	0,802
stosować	246	2	172/74	0,699
stosunek	187	3	93/75/19	0,497
strona	763	3	329/224/210	0,431
sztuka	232	3	175/29/28	0,754
światło	166	2	163/3	0,982
trafić	240	3	198/37/5	0,825
tworzyć	255	2	202/53	0,792
udać	337	3	288/45/4	0,855
uwaga	386	2	329/57	0,852
uważać	468	2	434/34	0,927
ważny	489	2	477/12	0,975
wiek	326	2	188/138	0,577
wolny	169	3	91/66/12	0,538
wpływ	230	2	208/22	0,904
wprowadzić	226	2	218/8	0,965
wydać	304	5	158/53/46/38/9	0,52
wydawać	420	5	300/72/34/12/2	0,714
wyglądać	287	3	279/6/2	0,972
wyjść	319	6	230/49/20/11/8/1	0,721

Słowo	Liczba wystąpień	Liczba sensów	Dystrybucja sensów	MFS
wynosić	127	2	111/16	0,874
wysokość	205	2	203/2	0,99
wystąpić	138	2	135/3	0,978
występować	204	3	139/63/2	0,681
zagrać	66	3	27/22/17	0,409
zająć	224	4	128/91/4/1	0,571
zajmować	254	3	170/78/6	0,669
zakład	327	2	316/11	0,966
zasada	351	2	350/1	0,997
zawierać	158	2	138/20	0,873
zawód	86	2	74/12	0,86
zawrzeć	164	5	86/69/6/2/1	0,524
zdawać	166	4	99/28/28/11	0,596
ziemia	312	3	167/124/21	0,535
złożyć	267	5	185/63/15/2/2	0,693
zostać	1 636	4	1 331/148/96/61	0,814
zwrócić	187	3	80/80/27	0,428
Statystyki dla wszystkich słów łącznie				
–	34 186	–	–	0,783

13

Narzędzia do anotacji jednostek nazewniczych

Agata Savary, Jakub Waszczuk

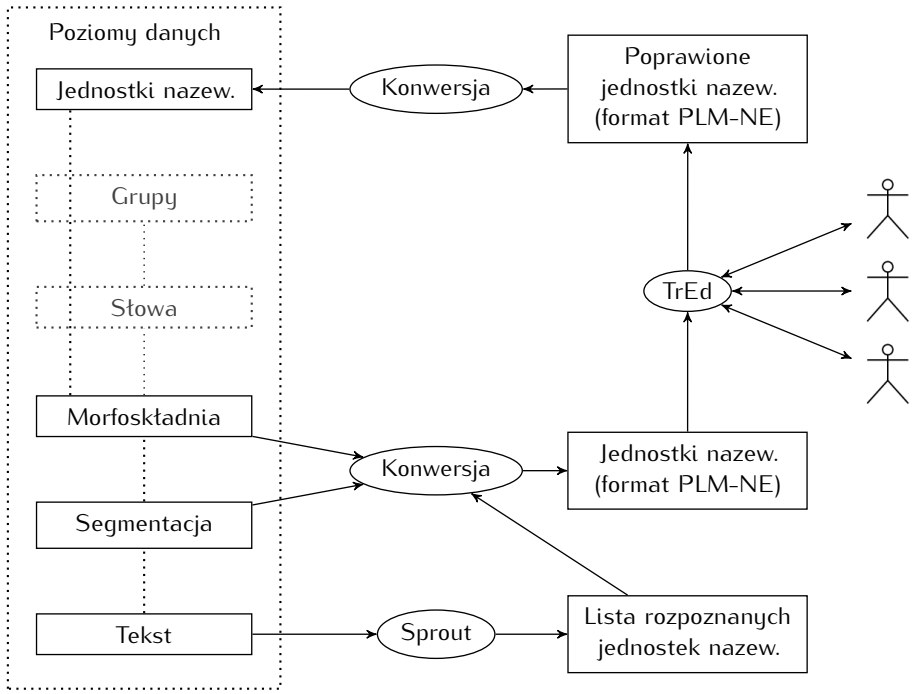
Anotacja lingwistyczna w NKJP wspierana była szeregiem narzędzi na wszystkich poziomach korpusu. W tym rozdziale prezentujemy trzy podstawowe narzędzia wykorzystywane do anotacji jednostek nazewniczych:

- *SProUT* – używany do wstępnej anotacji automatycznej podkorpusu milionowego w oparciu o zasoby leksykalne i gramatyki;
- *TrEd* – służący jako środowisko do ręcznej poprawy anotacji wykonanej SProUT-em;
- prototyp narzędzia statystycznego służącego do automatycznej anotacji pełnego korpusu.

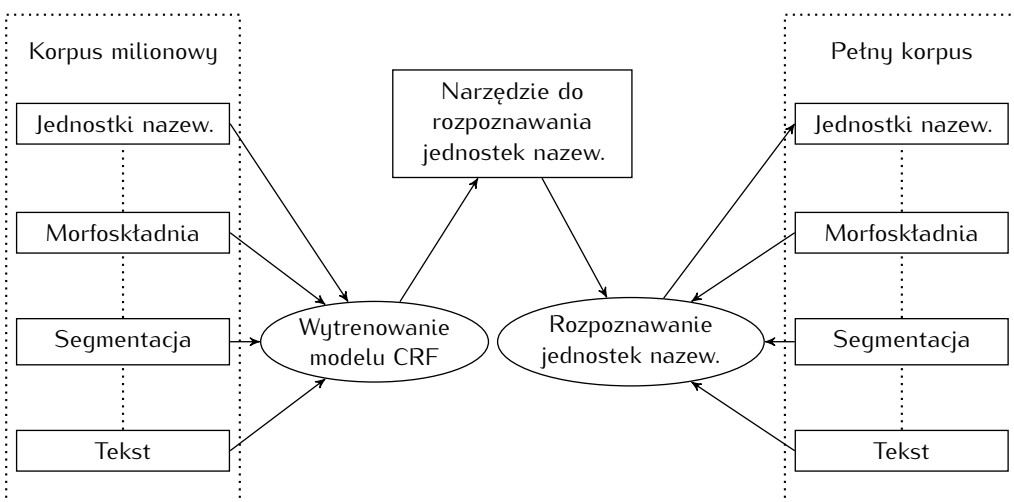
Rysunek 13.1 przedstawia przepływ danych podczas anotacji podkorpusu milionowego. „Czyste" teksty pochodzące z podkorpusu były najpierw przetwarzane przez system SProUT w celu automatycznej identyfikacji i klasyfikacji jednostek nazewniczych zgodnie z założeniami anotacji. Następnie wyniki SProUT-a były konwertowane na format narzędzia TrEd zwany PML. TrEd służył jako edytor drzew anotacyjnych – dzięki niemu dla każdego fragmentu tekstu dwaj niezależni anotatorzy mogli przeglądać, poprawiać i uzupełniać wyniki anotacji wstępnej, a tak zwani superanotatorzy – rozstrzygać niezgodności decyzji anotatorów. Po ukończeniu superanotacji plik wynikowy był konwertowany z formatu PML na ostateczny format zgodny ze standardem TEI P5.

Na rys. 13.2 pokazany jest przepływ danych w anotacji pełnego korpusu NKJP. Narzędzie statystyczne oparte na metodzie CRF (Conditional Random Fields; pol. warunkowe pola losowe; Wallach 2004, Sutton i McCallum 2007) tworzyło statystyczny model języka w oparciu o korpus milionowy, a dokładniej o jego poziom tekstowy, segmentacyjny, morfoskładniowy i poziom jednostek nazewniczych. Model ten był następnie stosowany do automatycznej anotacji pełnego korpusu.

Rysunek 13.1. Przepływ danych w procesie anotacji jednostek nazewniczych w korpusie milionowym



Rysunek 13.2. Przepływ danych w procesie anotacji jednostek nazewniczych w pełnym korpusie NKJP



13.1. Anotacja wstępna metodami regułowymi – platforma SProUT

SProUT (Becker i in. 2002, Drożdżyński i in. 2004) jest uniwersalną platformą do lingwistycznego przetwarzania tekstów. Narzędzie to ma kilka zalet istotnych z punktu widzenia przetwarzania języka polskiego: a) bogaty formalizm gramatyczny oferujący operatory wywodzące się z gramatyk regularnych, unifikację oraz przetwarzanie kaskadowe, b) moduły do szybkiego wyszukiwania haseł w słownikach dziedzinowych (ang. *gazetteers*), c) XML-owy format wyjściowy, zwany Sproutput, w postaci typowanych struktur cech, przy czym hierarchia typów może być zdefiniowana przez użytkownika. Dla wyboru SProUT-a jako narzędzia do anotacji wstępnej zasadniczy był fakt, że został on już wcześniej przystosowany do języka polskiego przez Piskorskiego i in. (2004), a ten sam autor stworzył gramatykę jednostek nazewniczych przeznaczoną do zadań ekstrakcji informacji (Piskorski 2005). Zasoby te musiały jednak zostać dostosowane do wymagań zadania anotacji jednostek nazewniczych w NKJP, co opisuje poniższy punkt p. 13.1.1. Dalsze szczegóły tego procesu opisane są przez Savary i Piskorskiego (2011).

13.1.1. Analiza i synteza morfologiczna

W analizie jednostek nazewniczych jedną z zasadniczych ról odgrywa analiza i synteza morfologiczna. Była ona realizowana dzięki analizatorowi *Morfeusz* (Woliński 2006), który stosuje bogaty zbiór znaczników zdefiniowanych według kryteriów morfologicznych i częściowo składniowych (Przepiórkowski i Woliński 2003). W NKJP używano dwóch wersji Morfeusza. Starsza z nich, zwana Morfeusz SIAT jest w pełni zintegrowana ze SProUT-em jako moduł do analizy morfologicznej języka polskiego. Rozpoznaje ona ok. 1,8 milionów form odmiany, wśród których znajduje się bardzo niewiele nazw własnych. Nowsza wersja, tzw. Morfeusz SGJP, rozszerzona o moduł syntezy (Savary i in. 2009), była używana jako samodzielne narzędzie do odmiany słowników dziedzinowych opisanych w następnym punkcie. Zawiera ona *Słownik gramatyczny języka polskiego* (Saloni i in. 2007) liczący ok. 4 miliony form odmiany odpowiadających 250 tysiącom form podstawowych. Wsród nich znajduje się ok. 10 tysięcy jednowyrazowych nazw własnych (głównie imion, nazwisk i nazw miejscowych).

13.1.2. Słownik dziedzinowy

Słowniki dziedzinowe (ang. *gazetteers*) pozwalają na dostosowanie platformy SProUT do danego zastosowania. Zawierają one listy form odmiany dla haseł,

specjalistycznych lub pochodzących z języka ogólnego, wraz z towarzyszącymi im atrybutami natury ontologicznej, morfologicznej i in. Odpowiedni moduł SProUT-a pozwala na szybkie wyszukiwanie tych form bezpośrednio w tekście i przypisywanie im pochodzących ze słownika atrybutów. Bogata morfologia języka polskiego sprawia, że słowniki takie mogą przybierać bardzo pokaźne rozmiary, dlatego konieczne są specjalne moduły do ich kompresji i dostępu w czasie rzeczywistym (Daciuk i Piskorski 2006, Budiscak i in. 2009).

W projekcie NKJP stworzony został słownik dziedzinowy zawierający ok. 290 tysięcy form odmiany odpowiadających ok. 55 tysiącom form podstawowych. Zgromadzono w nim:

1. słownik pochodzący z pracy Piskorskiego (2005), zawierający nieodmienione nazwy angielskie i niemieckie, ok. 60 tys. form odmiany polskich i obcojęzycznych nazwisk odpowiadających 1,5 tys. form podstawowych oraz niektóre słowa definiujące konteksty występowania jednostek nazewniczych (nazwy zawodów i funkcji, człony wyrażeń czasowych itp.);
2. przymiotniki relacyjne i nazwy mieszkańców polskich miejscowości pochodzące ze słownika Kubiak-Sokół i Łazińskiego (2007);
3. nazwy państw pochodzące ze źródeł Komisji Standaryzacji Nazw Geograficznych poza Granicami Rzeczypospolitej Polskiej[1];
4. wydobyte z Wikipedii[2] nazwy stolic administracyjnych państw, głównych rzek świata, regionów historycznych Europy, łańcuchów górskich oraz przymiotników i nazw obywateli państw świata;
5. nazwy 200 największych miast Polski pochodzące z *World Gazetteer*[3];
6. nazwiska polskie pochodzące z serwisu heraldycznego[4].

Hasła pochodzące ze wszystkich powyższych źródeł, za wyjątkiem pierwszego, były odmieniane za pomocą Morfeusza SGJP (zob. p. 13.1.1) – o ile oczywiście były one mu znane – a następnie przypisywano im typy (GTYPE), formy podstawowe (G_LEMMA), wartości kategorii morfologicznych (G_NUMBER, G_CASE, G_GENDER), ewentualne typy derywacji (G_DERIV_TYPE) i bazy derywacyjne (G_DERIVED_FROM), jak również identyfikatory wskazujące na pochodzenie zarówno samych haseł (G_SOURCE) jak i ich odmiany (G_INFL_SOURCE). Pewne nieliczne nazwy wielowyrazowe opisane i odmienione zostały narzędziem *Multiflex* (Savary i in. 2009). Przykłady (13.1)–(13.3) prezentują ostateczną postać haseł słownika dziedzinowego, a na rys. 13.3 podano jego szczegółową zawartość[5].

[1] `http://www.gugik.gov.pl/komisja/`.

[2] `http://pl.wikipedia.org/wiki`.

[3] `http://www.world-gazetteer.com`.

[4] `http://www.futrega.org/etc/nazwiska.html`.

[5] Proporcje ilościowe między formami podstawowymi a formami odmiany wydają się zaskakujące dla niektórych kategorii, jak np. imiona czy organizacje. Wynika to z faktu, że część haseł ze

(13.1) **Helskim** | GTYPE:gaz_city_deriv | G_LEMMA:helski | G_NUMBER:singular | G_CASE:loc | G_GENDER:masc1_masc2_masc3_neutrum1_neutrum2 | G_DERIV_TYPE:reladj | G_DERIVED_FROM:Hel | G_SOURCE:PWN_Miejscowe | G_INFL_SOURCE:Morfeusz | G_LETTER_CASE:first-upper

(13.2) **Skarżyskiem-Kamienną** | GTYPE:gaz_city | G_LEMMA:Skarżysko-Kamienna | G_GNUMBER:singular | G_CASE:ins | G_GENDER:neutrum2 | G_SOURCE:WorldGazetteer | G_INFL_SOURCE:Morfeusz_Multiflex

(13.3) **inspektorowi** | GTYPE:gaz_position | G_LEMMA:inspektor | G_CASE:dat | G_GENDER:masc1 | G_GNUMBER:singular

Rysunek 13.3. Zawartość polskiego słownika dziedzinowego i gramatyki dla SProUT-a, wykorzystywanych w NKJP

Kategoria	Formy podstawowe	Formy odmiany
Imiona	17 068	39 461
Nazwiska	17 474	85 651
Organizacje	1 752	1 863
Państwa i regiony	373	3 472
Miasta	2 952	6 569
Rzeki i góry	361	2 003
Przymiotniki	1 871	128 424
Mieszkańcy	12 292	19 387
Słowa kontekstowe	564	2 409
RAZEM	54 707	289 239

Typy	Reguły
Osoby	29
Organizacje	20
Nazwy miejscowe	25
Wyrażenia czasu	24
Derywacje	5
Różne	17
RAZEM	120

13.1.3. Hierarchia typów i gramatyka

W projekcie NKJP wykorzystana została gramatyka jednostek nazewniczych stworzona przez Piskorskiego (2005), zawierająca ok. 160 reguł. Była ona pierwotnie przeznaczona do ekstrakcji informacji, dlatego musiała zostać dostosowana

słownika Piskorskiego, jak i część przymiotników, nie jest odmieniona, a jedynie podana w formie podstawowej.

do zadania anotacji według reguł opisanych w rozdz. 9. Modyfikacji poddano w szczególności hierarchię typów:

1. Wprowadzono taksonomię przedstawioną na rys. 9.2, której węzły stanowią wartości atrybutu NE_TYPE.
2. Uzupełniono atrybuty haseł występujących w słowniku dziedzinowym, np. G_DERIV_TYPE, G_DERIVED_FROM, G_SOURCE, G_INFL_SOURCE, por. przykłady (13.1)–(13.3).
3. Wprowadzono 170 leksemów typowych dla kontekstów występowania nazw, czyli tzw. *dowodów wewnętrznych* i *zewnętrznych*, np. *Góra*, *Półwysep*, *rzeka*, *zatoka*.
4. Zdefiniowano tekstowy atrybut TREE, mający zasadnicze znaczenie przy opisywaniu nazw zagnieżdżonych.

Każda reguła gramatyki SProUT-a składa się ze strony lewej, pozwalającej na wyszukiwanie wzorców w tekście i przypisywanie wartości atrybutom, oraz prawej, tworzącej struktury wynikowe. Rysunek 13.4 przedstawia regułę o nazwie `city_reladj_gaz_based` pozwalającą na rozpoznanie w tekście przymiotnika relacyjnego występującego w słowniku dziedzinowym, np. *Helskim* z przykładu (13.1). Atrybuty GTYPE i G_DERIV_TYPE mają wymagane wartości stałe `gaz_city_deriv` i `reladj`, zaś pozostałe atrybuty są zmiennymi (ich identyfikatory zaczynają się znakiem #). Wartości atrybutów SURFACE, G_LEMMA, G_NUMBER, G_CASE, G_GENDER, i G_DERIVED_FROM są przepisywane bezpośrednio ze słownika dziedzinowego. Atrybuty CSTART i CEND, oznaczające pozycję znaku rozpoczynającego i kończącego znalezioną nazwę, są automatycznie inicjalizowane przez analizator. Po prawej stronie tej reguły niektóre atrybuty są przepisywane ze strony lewej, a niektóre mają wartości stałe. W szczególności atrybut TREE jest konstruowany przez *operator funkcyjny* `ConcWithBlanks`, pozwalający na połączenie wybranych atrybutów, tak aby stworzona została następująca struktura (przy założeniu, że słowo *Helskim* wystąpiło na pozycji 11 345): `[ Helskim | Helski | settlement | relAdj | Hel | 11345 | 11351 | prio_2 ]`. Informacja `prio_2` opisuje priorytet danej interpretacji w stosunku do innych ewentualnie z nią konkurujących, mogących dostarczać mniej lub bardziej poprawnych atrybutów (np. w przypadku, gdy słowo rozpoznane jest wyłącznie dzięki kryterium pisowni wielką literą, lecz bez możliwości jego lematyzacji).

Reguły gramatyczne mogą zawierać odwołania do innych reguł (poprzez dyrektywę `@seek`). Na przykład rys. 13.5 prezentuje regułę `geogr_names_int_proof_1`, która wywołuje dwie reguły, w tym `city_reladj_gaz_based` omawianą powyżej. Pozwala to na rozpoznanie ciągu złożonego z rzeczownika pisanego wielką literą zaczerpniętego z listy dowodów wewnętrznych (np. *Półwyspie*) oraz przymiotnika relacyjnego (np. *Helskim*). Oba człony muszą występować w zgodnej

Rysunek 13.4. Reguła gramatyczna rozpoznająca przymiotnik relacyjny pochodzący od nazwy miejscowości i znajdujący się w słowniku dziedzinowym

```
city_reladj_gaz_based :/
 gazetteer & [SURFACE #surface, G_LEMMA #lemma, GTYPE gaz_city_deriv,
              G_NUMBER #number, G_CASE #case, G_GENDER #gender,
              G_DERIV_TYPE reladj, G_DERIVED_FROM #df, CSTART #s, CEND #e]
->
ne-nkjp & [SURFACE #surface, BASE #lemma, NE_TYPE settlement,
           MORPH agr-nkjp & [NE_NUMBER #number, NE_CASE #case,
                             NE_GENDER #gender],
           TREE #tree, DERIV_TYPE reladj, DERIVED_FROM #df,
           CSTART #s, CEND #e],
where #tree=ConcWithBlanks("[", #surface, "|", #lemma,
                           "| settlement | relAdj |",
                           #df, "|",  #s, "|", #e, "| prio_2 ]").
```

Rysunek 13.5. Reguła gramatyczna rozpoznająca nazwy geograficzne zawierające dowód wewnętrzny, takie jak *Półwysep Helski*

```
geogr_names_int_proof_1 :/
   @seek(capitalized_noun) &
   [SURFACE #surface1, STEM int_proof_geog_name & #lemma1,
    INFL infl_noun & [GENDER_NOUN #g, NUMBER_NOUN #n, CASE_NOUN #c],
    CSTART #s1, CEND #e1]
   @seek(city_reladj_gaz_based) &
   [SURFACE #surface2, BASE #lemma2,
    MORPH agr-nkjp & [NE_NUMBER #n, NE_CASE #c, NE_GENDER #g],
    TREE #tree2, CSTART #s2, CEND #e2]
->
ne-nkjp & [SURFACE #surface, BASE #lemma, NE_TYPE geog_name, TREE #tree,
           CSTART #s1, CEND #e2],
where Capitalized(#surface2),
      #surface=ConcWithBlanks(#surface1, #surface2),
      #lemma=ConcWithBlanks(#lemma1, #lemma2),
      #tree=ConcWithBlanks(#tree2, "[", #surface, "|", #lemma,
                           "| geogName |", #s1, "|", #e2, "| prio_1 ]").
```

liczbie, przypadku i rodzaju, co wyrażone jest przez wspólne zmienne unifikacyjne `#n`, `#c` i `#g`. Struktura wynikowa tak rozpoznanej nazwy przedstawiona jest na rys. 13.6. Zauważmy, że atrybut TREE zawiera tu informacje o nazwie nadrzędnej, jak i zagnieżdżonej. Atrybut ten jest przekształcany na drzewo anotacyjne podczas konwersji na format PML (zob. p. 13.2).

Dolna tabela z rys. 13.3 podaje ilość reguł używanych w anotacji wstępnej korpusu milionowego NKJP w rozbiciu na typy nazw.

Rysunek 13.6. Struktura wynikowa nazwy *na Półwyspie Helskim* rozpoznanej dzięki regule `geogr_names_int_proof_1`

SURFACE	Półwyspie Helskim
BASE	Półwysep Helski
TYPE	geogName
TREE	[Helskim | Helski | settlement | relAdj | Hel | 11345 | 11351 | prio_2]
	[Półwyspie Helskim | Półwysep Helski | geogName | 11335 | 11351 | prio_1]
CSTART	11335
CEND	11351

13.1.4. Wyniki

Szczegółowa ewaluacja jakości wyników SProUT-a w anotacji jednostek nazewniczych w NKJP została opisana w artykule Savary i Piskorskiego (2011). Została ona zrealizowana na podzbiorze korpusu zawierającym ok. 56 tys. jednostek nazewniczych, dla którego anotacja ręczna wykonana została po ustabilizowaniu zasobów i gramatyk SProUT-a. Eksperymenty zaprojektowano według dwóch kryteriów: a) wzięcie pod uwagę wszystkich jednostek, w tym zagnieżdżonych, lub też wyłącznie jednostek o maksymalnej długości (odpowiadających korzeniom drzew anotacyjnych), b) analiza wszystkich atrybutów bądź wyłącznie typów i podtypów. Eksperymenty te zaowocowały zatem czterema zestawami wyników, które pokazują, że:

1. Ogólna dokładność (ang. *precision*) waha się w granicach od 68% do 78%, a pełność (ang. *recall*) od 35% do 39%.
2. Wyniki są z oczywistych względów niższe, gdy brany jest pod uwagę pełen zestaw atrybutów, a nie tylko typy i podtypy nazw; różnice w tych dwóch scenariuszach mieszczą się w granicach od 2% do 13% dokładności i od 2% do 5% pełności.
3. Najlepsze wyniki uzyskiwane są dla wyrażeń czasowych, a najgorsze dla nazw organizacji.

Scenariusz, w którym brane są pod uwagę wszystkie jednostki i wszystkie atrybuty, najściślej odpowiada użyciu SProUT-a w anotacji wstępnej NKJP. Uzyskana dla tego scenariusza dokładność w wysokości 72% oznacza, że taki właśnie procent wszystkich anotacji zaproponowanych przez SProUT-a nie wymaga żadnej poprawki ze strony anotatora. Pełność sięgająca 36% dla tego samego scenariusza wskazuje na procent wszystkich istniejących w korpusie jednostek nazewniczych, które oznaczone zostają całkowicie poprawnie przez to narzędzie. Jego użycie stanowi zatem znaczne usprawnienie anotacji ręcznej. We wspomnianym artykule znaleźć można również szczegółową analizę błędów popełnianych przez SProUT-a.

Wspomnijmy na koniec, że czas przetwarzania tym narzędziem korpusu zawierającego 1 milion słów wyniósł 45 minut i 18 sekund. Wynik ten można uznać za niedostateczny dla przetwarzania dużych ilości tekstów w czasie rzeczywistym, jednak jest on zadowalający w kontekście anotacji wstępnej wykonywanej offline.

13.2. Ręczna poprawa anotacji podkorpusu milionowego – platforma TrEd

Ręczna poprawa i uzupełnianie wstępnej anotacji automatycznej były najbardziej pracochłonnymi etapami na każdym poziomie NKJP. Anotatorzy mieli bowiem za zadanie przejrzenie i poprawę anotacji dokonanej SProUT-em dla każdego zdania w korpusie w celu: a) zatwierdzenia poprawnych drzew anotacyjnych (ang. *true positives*), b) usunięcia lub poprawy błędnych drzew (ang. *false positives*), c) dodania drzew niedostrzeżonych (ang. *false negatives*)[6]. Opracowanie optymalnych narzędzi przeznaczonych do tych zadań miało zasadniczy wpływ na czas trwania projektu oraz jakość wyników. Dokonaliśmy ewaluacji kilku platform anotacyjnych – głównie Synpathy[7], MMAX[8] oraz GATE (Wilcock 2009) – zanim wybór nasz padł na edytor drzew TrEd[9] (Pajas i Štěpánek 2008) stworzony w ramach projektu anotacji Praskiego Korpusu Zależnościowego (Prague Dependency Treebank, PDT, Böhmová i in. 2003). Ma on wiele zalet potrzebnych w projekcie NKJP: 1. możliwość przetwarzania korpusów anotowanych wcześniej metodami zewnętrznymi, 2. dopuszczanie zewnętrznej (ang. *stand-off*) anotacji wielopoziomowej, 3. otwarty XML-owy format danych zwany PML, łatwo przystosowywalny za pomocą tzw. *rozszerzeń* do różnorodnych zastosowań (m.in. dostosowanie edycji drzew zależnościowych do drzew składnikowych nie nastręczyło wielu trudności), 4. łatwa obsługa anotacji w formie drzew, 5. ergonomiczny interfejs, dostosowywalny do indywidualnych potrzeb poprzez arkusze stylów, 6. możliwość równoległej edycji dwóch konkurencyjnych anotacji, 7. bogata dokumentacja, 8. techniczna niezawodność. W NKJP TrEd był używany do anotacji zarówno jednostek nazewniczych jak i grup składniowych.

[6] Szczegółowa analiza błędów popełnianych przez SProUT-a znajduje się w publikacji Savary i Piskorskiego (2011).

[7] `http://www.lat-mpi.eu/tools/synpathy`.

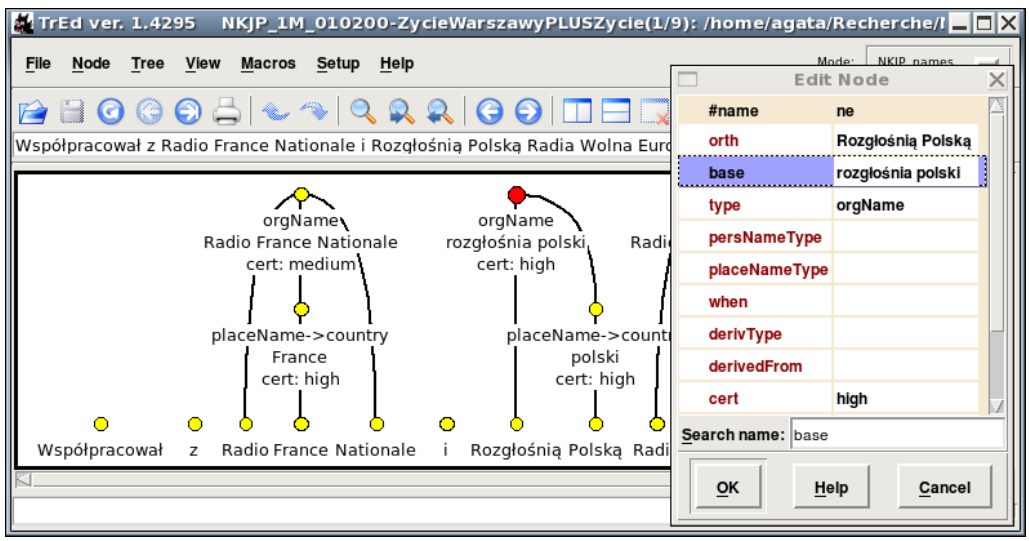
[8] `http://mmax2.sourceforge.net`.

[9] `http://ufal.mff.cuni.cz/~pajas/tred/`.

13.2.1. Warsztat anotatora

Na rys. 13.7 pokazane jest okno robocze TrEda w trakcie ręcznej poprawy anotacji zdania *Współpracował z Radio France Nationale i Rozgłośnią Polską Radia Wolna Europa*. Zgodnie ze stworzonymi na potrzeby NKJP regułami wizualizacji, na najniższym poziomie znajdują się węzły reprezentujące segmenty poziomu morfoskładniowego. Wyższe węzły reprezentują możliwie zagnieżdżone jednostki nazewnicze, np. *France* jest nazwą zagnieżdżoną w *Radio France Nationale*. Dzięki wbudowanym w TrEda skrótom klawiszowym można łatwo dodawać i usuwać węzły oraz krawędzie, a także nawigować między zdaniami i plikami. Jednoczesny wybór kilku węzłów, dodawanie węzłów określonego typu i szybka zmiana wartości atrybutów umożliwione zostały przez rozszerzenia stworzone specjalnie w ramach NKJP.

Rysunek 13.7. Okno robocze TrEda w trakcie poprawy formy podstawowej nazwy *Rozgłośnia Polska*. Formy podstawowe członów nazwy (*rozgłośnia polski*) są skopiowane z poziomu morfoskładniowego



Każdej jednostce towarzyszą jej atrybuty: forma podstawowa (`@base`) lub znormalizowana (`@when`), typ (`@type`) i ewentualny podtyp (`@persNameType` lub `@placeNameType`), ewentualny typ derywacji (`@derivType`) i baza derywacyjna (`@derivedFrom`) oraz stopień pewności anotacji (`@cert`) z komentarzem (`@certComment`) w przypadku stopnia `medium` or `low`. Szczegółowe reguły przypisywania atrybutów oraz problemy z tym związane opisane są w rozdz. 9. Edycja atrybutów możliwa jest w osobnym oknie otwieranym po kliknięciu na dany węzeł lub na pojedynczy atrybut. Atrybuty `@base`, `@derivedFrom` i `@certComment`

są w pełni edytowalne, pozostałe atrybuty zaś wybierane są z list, co ogranicza ryzyko błędów. Zamykanie danego pliku automatycznie uruchamia skrypt sprawdzający istnienie i spójność wymaganych atrybutów.

Po ukończeniu superanotacji opisanej w następnym punkcie każdy plik jest konwertowany z formatu PML na ostateczny format TEI P5, jak to zostało opisane w rozdz. 10. W procesie tym zachodzi m.in. przekształcanie atrybutów PML na elementy ⟨f⟩, np. atrybut @base staje się ⟨f name="base"⟩, podobnie jest dla @when, @orth, @type, @derivType i @derivedFrom. Atrybuty @cert i @certComment zmieniają przy tym nazwy na ⟨f name="certainty"⟩ i ⟨f name="comment"⟩, atrybuty @persNameType i @placeNameType są zaś grupowane w jeden element ⟨f name="subtype"⟩.

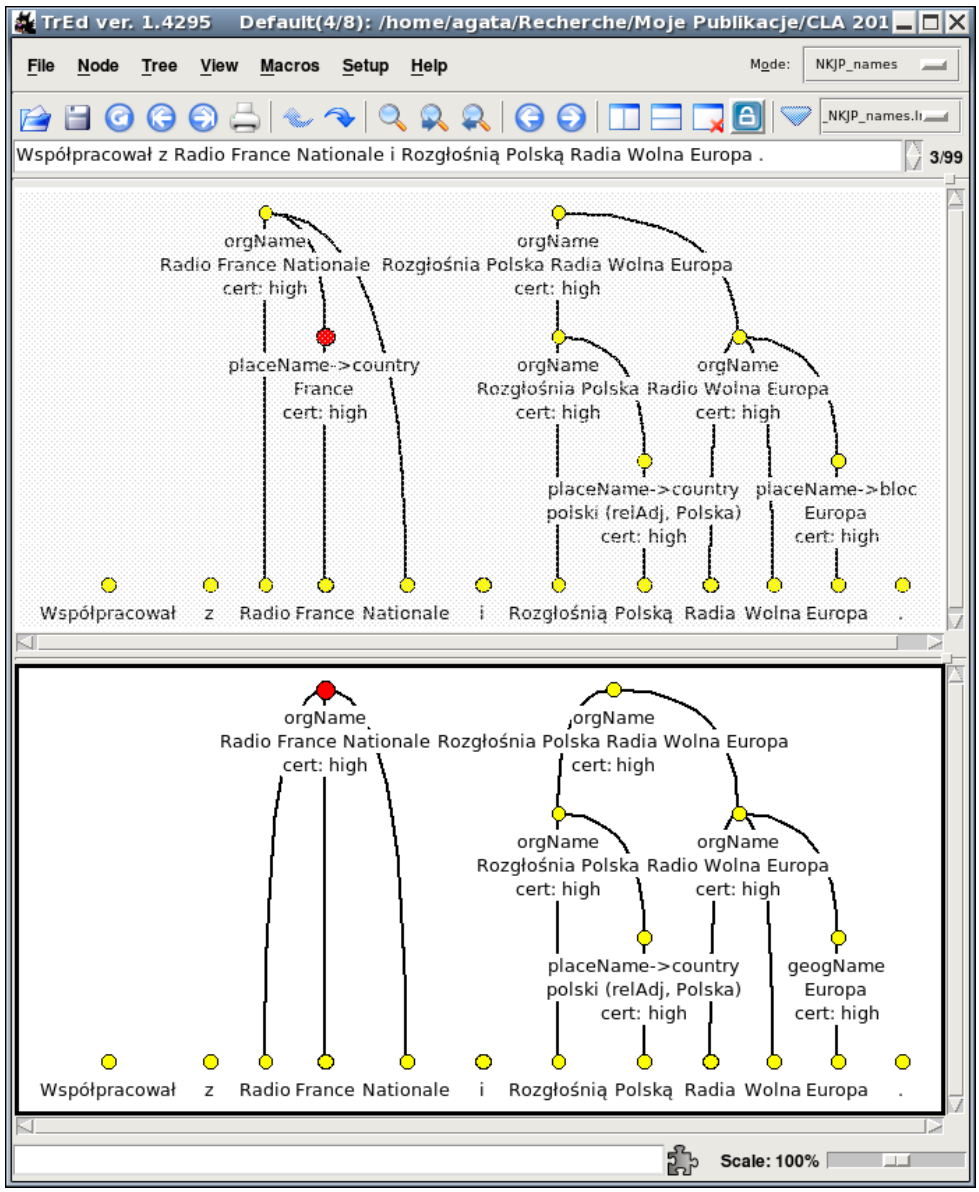
13.2.2. Warsztat superanotatora

Jak wspominano, każdy tekst korpusu poprawiany był ręcznie przez dwóch niezależnych anotatorów, a następnie niezgodności ich ocen były rozstrzygane przez tzw. superanotatora. W celu zwiększenia obiektywności osądów, trzymano się przy tym następujących zasad: 1. anotatorzy tego samego tekstu nie znali nawzajem swoich anotacji, jedyne uzgadnianie ocen możliwe było poprzez listę dyskusyjną, 2. w roli superanotatora wystąpić mogły osoby o poprzednim bogatym doświadczeniu w roli anotatora, 3. nie można było superanotować tekstu, który się wcześniej anotowało.

Rozwinięte w ramach NKJP rozszerzenia TrEda pozwalały na automatyczne wyszukiwanie rozbieżności między dwiema anotacjami, jak również na automatyczne przenoszenie fragmentów anotacji. Rysunek 13.8 pokazuje okno TrEda w trakcie superanotacji tego samego zdania, co na rys. 13.7. Załóżmy, że niższe okno, odpowiadające pierwotnie wynikom pracy anotatora *a2*, zostało wybrane przez superanotatora jako to, którego zawartość przekształcana będzie w celu uzyskania ostatecznej wersji anotacji. W wyższym oknie, odpowiadającym anotatorowi *a1*, widnieje węzeł dla nazwy kraju *France*, która nie została opisana przez *a2*. Węzły odpowiadające tej rozbieżności zostały podświetlone na czerwowo. Pojedynczy skrót klawiszowy pozwala w tej sytuacji na wstawienie brakującego węzła do niższego okna, powyżej węzła *France* a poniżej *Radio France Nationale*, tak iż pozostałe węzły nie ulegają zmianie.

Automatyczne wykrywanie i przenoszenie różnic w anotacji dotyczy nie tylko brakujących lub zbędnych węzłów, ale również ich atrybutów. W sytuacji przedstawionej na rys. 13.7 następną wykrytą różnicą będzie typ nazwy *Europa* (placeName→bloc dla *a1* i geogName dla *a2*). Ponieważ jednak poprawny typ został wybrany przez *a2*, propozycja pochodząca od *a1* nie zostanie przeniesiona do niższego okna.

Rysunek 13.8. Okno robocze TrEda w trakcie superanotacji. Węzły podświetlone na czerwono zostały znalezione automatycznie jako zawierające niezgodności między dwoma anotatorami



13.2.3. Zarządzanie plikami

Anotacja i superanotacja dokonywane według wyżej opisanych reguł wymagały efektywnego systemu pobierania, odsyłania i archiwizowania anotowa-

nych plików. Funkcję tę pełniło dostosowane do naszych potrzeb repozytorium svn[10] oraz XML-owa baza danych przechowująca informacje o pobieranych i odsyłanych plikach. Każdy anotator i superanotator łączył się z serwerem *svn* wyłącznie w momencie pobierania i odsyłania określonej przez siebie i globalnie ograniczonej liczby plików, a przez pozostały czas pracował w trybie offline. Miał on przyznane wyłącznie minimalne prawa do modyfikowania zawartości repozytorium, a mianowicie do modyfikacji przyznanych mu plików. Przyznawanie plików anotatorom oraz przenoszenie ich między katalogami w repozytorium wykonywane było automatycznie przez skrypty. Dodatkowe procedury pozwalały również na zapewnienie spójności repozytorium oraz bazy danych – m.in. dwaj użytkownicy nie mogli komunikować się z repozytorium w tym samym czasie. XML-owa baza danych pozwalała na szybkie wyszukiwanie informacji o plikach według kilku kryteriów: identyfikator (super)anotatora obsługującego plik, data pobrania lub odesłania, ilość zdań i słów oraz status (nieprzyznany, pobrany, ukończony itd.).

13.2.4. Zgodność anotatorów

Zgodnością anotatorów (ang. *inter-annotator agreement*) nazywamy intuicyjnie stopień, w jakim anotacje tego samego tekstu wykonane przez dwie osoby pokrywają się ze sobą. Jest to klasyczny wskaźnik jakości dokumentacji i metodologii projektu anotacyjnego, co ściśle związane jest z przydatnością przedstawionego powyżej warsztatu anotatora i superanotatora, jak również reguł anotacji opisanych w rozdz. 9.

Mimo swojej intuicyjnej jasności, wskaźnik zgodności anotatorów jest w naszym projekcie pojęciem nieco skomplikowanym. Powszechnie stosuje się tzw. miarę *kappa* (Cohen 1960), biorącą pod uwagę zarówno procent zgodnych decyzji obu anotatorów, jak i prawdopodobieństwo, że zgodność ta nastąpiła przypadkiem. Niestety miarę kappa można stosować tylko wówczas, gdy znamy anotowane próbki. Innymi słowy, musielibyśmy z góry wiedzieć, gdzie w tekście występują jednostki nazewnicze, a następnie sprawdzać, jak zostały one opisane przez anotatorów. O problemie tym piszą Bejček i Straňák (2010). Zaproponowany w tej publikacji sposób szacowania zgodności anotatorów wydaje się przekonujący, jednak nieco skomplikowany.

[10] *Subversion* – system kontroli wersji, pozwalający m.in. na archiwizację zmian zawartych w nim plików.

Na potrzeby NKJP oszacowaliśmy zgodność anotatorów w sposób prostszy, opierając się na klasycznych pojęciach z dziedziny ekstrakcji informacji: dokładności (ang. *precision*) i pełności (ang. *recall*). Chodziło o wyliczenie tych dwóch miar, wraz z ich średnią harmoniczną (ang. *f-measure*), dla jednego anotatora względem drugiego. Przedtem należało jednak jasno zdefiniować, kiedy daną jednostkę uważaliśmy za opisaną poprawnie, a kiedy nie. Specyfikę tego rachunku stanowi fakt, że jednostki nazewnicze dekorowane były drzewami anotacyjnymi, z możliwymi wielopoziomowymi zagnieżdżeniami, jak również atrybutami zależnymi od ich typów głównych. Dlatego każdy węzeł danego korpusu, zarówno nadrzędny jak i zagnieżdżony, uznawaliśmy za poprawny, jeżeli odpowiadał on węzłowi w korpusie drugiego anotatora:

1. mającemu te same wartości atrybutów,
2. zawierającemu w swoim drzewie te same segmenty poziomu morfoskładniowego.

Wyniki obliczeń tak zdefiniowanej zgodności anotatorów przedstawione są w tab. 13.1. *Nazwy osobowe* odpowiadają typowi `persName` i jego 3 podtytpom (zob. rozdz. 9), *nazwy miejscowe* obejmują typy `geogName` i `placeName` (wraz z jego 5 podtypami). *Nazwy organizacji* odnoszą się do typu `orgName`, a *wyrażenia czasowe* do `date` i `time`. Wreszcie *derywacje* oznaczają jednostki mające atrybuty `@derivType` równe `relAdj` lub `persDeriv` i pochodzące od nazw dowolnego typu i podtypu.

Tabela 13.1. Zgodność anotatorów jednostek nazewniczych

Nazwy osobowe	**Nazwy miejscowe**	**Nazwy organizacji**	**Wyrażenia czasowe**	**Derywacje**	**RAZEM**
0,89	0,78	0,69	0,88	0,71	0,83

Wyniki te wskazują, iż zdefiniowana w rozdz. 9 anotacja jednostek nazewniczych jest zadaniem stosunkowo trudnym. Można to częściowo uzasadnić poprzez fakt, iż większość uwzględnianych tu atrybutów jest natury semantycznej. Kategoriami najtrudniejszymi do oznaczania okazują się nazwy organizacji oraz derywacje. Najlepsze wyniki uzyskiwane są zaś dla nazw osobowych. Jedną z przyczyn tego stanu rzeczy jest prawdopodobnie rzadsze podleganie przez te nazwy zjawiskom takim jak metonimia (por. p. 9.4.1).

Bardziej szczegółowa definicja wskaźnika zgodności anotatorów znajduje sie w pracy Waszczuka i in. (2011).

13.3. Anotacja pełnego korpusu przy użyciu uczenia maszynowego

13.3.1. Anotacja jednostek nazewniczych jako problem etykietowania

Do zaanotowania pełnego korpusu wybrana została metoda *Joined Label Tagging* opisana w artykule Alex i in. (2007). W metodzie tej problem rozpoznawania jednostek nazewniczych sprowadzony jest do problemu etykietowania (tagowania) przy użyciu konwencjonalnego tagera sekwencyjnego. Przez tager sekwencyjny rozumiemy tutaj dowolne narzędzie, które przyporządkowuje segmentom etykiety (tagi) z pewnego z góry zadanego zbioru. Zakładamy również, że narzędzie to można „nauczyć" rozpoznawania etykiet w oparciu o zaanotowany ręcznie jednostkami nazewniczymi korpus milionowy. W literaturze można znaleźć również rozwiązania, w których drzewiaste struktury jednostek modelowane są w sposób bezpośredni (zob. np. Finkel i Manning 2009). Metoda Joined Label Tagging jest jednak dużo prostsza do zaimplementowania i pozwala na wykorzystanie jednego z wielu istniejących narzędzi do tagowania sekwencyjnego. Użycie tej metody odbywa się jednak kosztem mniej intuicyjnej reprezentacji jednostek. Ponadto za pomocą etykiet trudno jest wyrazić niektóre relacje zachodzące pomiędzy jednostkami (np. zagnieżdżanie). Na początek zajmiemy się problemami związanymi z samą reprezentacją jednostek za pomocą etykiet, a następnie opiszemy szczegóły wykorzystanego tutaj tagera sekwencyjnego.

Reprezentacja jednostek przy użyciu etykiet

W jaki sposób wykorzystać tager do rozpoznawania jednostek nazewniczych? Należy najpierw otagować zdanie sekwencją etykiet, a następnie odtworzyć jednostki zakodowane w postaci tej sekwencji. Potrzebujemy w takim razie metodę kodowania jednostek za pomocą sekwencji etykiet oraz odwrotnie – metodę dekodowania sekwencji do zbioru jednostek. Wtedy po otagowaniu zdania etykietami będziemy w stanie odtworzyć zbiór jednostek, jeśli tylko sekwencja etykiet jest poprawna. Powiemy, że sekwencja etykiet jest poprawna, jeśli istnieje taki zbiór jednostek (określonych względem danego zdania), że sekwencja jest wynikiem jego zakodowania. Wprowadzamy pojęcie poprawności sekwencji, ponieważ w ogólnym przypadku tager nie gwarantuje, że wynikowa sekwencja jest poprawna.

Na początek rozważmy jedynie ciągłe, niezagnieżdżone oraz nienachodzące na siebie jednostki[11]. W tej sytuacji do reprezentacji jednostek można użyć standardowej i powszechnie znanej metody o nazwie *IOB encoding* (Ramshaw i Marcus 1995). Zgodnie z nią, początek jednostki typu `T` reprezentujemy etykietą `B-T`, kolejnym segmentom należącym do jednostki odpowiadają etykiety `I-T`, natomiast segmenty, które nie wchodzą w obręb żadnej jednostki, otrzymują etykietę `O`. Przykładowo, struktura jednostek dla frazy

(13.4) *[Czarny Staw]$_{geogName}$ leżący w [Tatrach Wysokich]$_{geogName}$*

po zakodowaniu przyjmie postać następującej sekwencji etykiet:

(13.5) *[B-geogName, I-geogName, O, O, B-geogName, I-geogName]*

natomiast fraza:

(13.6) *w Paryżu dnia 21 września 1960 r.*

z przykładu (9.5) zostanie zakodowana jako:

(13.7) *[O, B-settlement, B-date, I-date, I-date, I-date, I-date]*

Kodowanie takie jest jednoznaczne: a) zbiorowi jednostek występujących w zdaniu odpowiada dokładnie jedna sekwencja etykiet, b) poprawnej sekwencji etykiet odpowiada dokładnie jeden zbiór jednostek (w którym każda jednostka reprezentowana jest przez zbiór pokrywanych segmentów oraz swoją anotację, bez uwzględnienia form podstawowych i baz derywacyjnych). Przykładem niepoprawnej sekwencji jest ciąg, w którym etykieta typu `I-T` występuje bezpośrednio po etykiecie `O`. Jest to sekwencja, której nie można uzyskać przez zakodowanie zbioru rozłącznych, ciągłych i niezagnieżdżonych jednostek.

Na poziomie etykiet zapisujemy informacje nie tylko o typach (`type`) poszczególnych jednostek, ale również o podtypach (`subtype`) oraz rodzaju derywacji (`derivType`) – jeśli taki przypadek zachodzi. Od tej pory przyjmiemy rozszerzoną definicję *typu jednostki* – jest to wartość powstała przez sklejenie powyższych trzech informacji. Podtyp determinuje typ, dlatego w podawanych poniżej przykładach sekwencji etykiet typ często jest pomijany. Na poziomie etykiet nie będziemy natomiast reprezentować informacji o formie podstawowej (`base`) lub znormalizowanej (`when`) jednostki, a także o bazie derywacyjnej (`derivedFrom`). Są to cechy o potencjalnie nieograniczonym zbiorze wartości, dlatego ich rozpoznawanie przy użyciu opisywanej tutaj metody byłoby karkołomnym zadaniem.

[11] Przyjmujemy tutaj definicję, że jednostki nachodzą na siebie, jeśli segment jednej jednostki wchodzi w obręb drugiej lub jest położony między segmentami drugiej jednostki, jak w przykładach (9.13) i (9.21).

Jednostki nieciągłe

W sytuacji, gdy w zdaniu występują jednostki nieciągłe, można przypisać etykietę `O` segmentom położonym pomiędzy dwiema częściami jednostki, ale nie należącym do tej jednostki. Tak więc przedstawioną w przykładzie (9.23) nazwę osobową:

(13.8) *Gisela Froemel, z domu Kopka*

można, po pominięciu podtypów i zagnieżdżania, zakodować za pomocą następującej sekwencji etykiet:

(13.9) *[B-persName, I-persName, O, O, O, I-persName]*

Ponieważ założenie o niezachodzeniu na siebie jednostek ciągle jest w mocy, kodowanie również w tym wypadku jest jednoznaczne, jak również dekodowanie poprawnych sekwencji etykiet (chociaż tutaj zbiór poprawnych sekwencji etykiet jest nieco inny niż poprzednio).

Jednostki zagnieżdżone

Rozważmy teraz zagnieżdżanie jednostek nazewniczych, które w korpusie zdarza się bardzo często, bowiem około 35% wszystkich jednostek występujących w korpusie milionowym stanowią jednostki zagnieżdżone. Kodowanie zbioru jednostek nazewniczych w sytuacji, gdy występuje zagnieżdżanie, można zrealizować przy użyciu jednej z następujących technik:

1. Osobno rozpoznajemy jednostki na poszczególnych poziomach zagnieżdżenia, przy czym przyjmujemy pewien maksymalny poziom zagnieżdżenia r. W ramach tego rozwiązania możemy również przyjąć kolejność rozpoznawania poszczególnych warstw jednostek:
   a) Najpierw rozpoznajemy najbardziej zagnieżdżone (najkrótsze) jednostki, następnie jednostki o jeden poziom wyżej itd., aż do osiągnięcia pustej warstwy lub poziomu zagnieżdżenia równego r. W trakcie rozpoznawania jednostek w i-tej warstwie możemy korzystać z informacji o jednostkach rozpoznanych na niższych poziomach.
   b) Rozpoznawanie zaczynamy od zewnętrznej warstwy i kontynuujemy, jak powyżej, do osiągnięcia warstwy pustej lub poziomu zagnieżdżenia równego r.
2. Dla danego segmentu, etykiety ze wszystkich poziomów zagnieżdżenia łączymy w jedną, złożoną etykietę. Przykładowo struktura jednostek dla frazy *Polska Akademia Nauk*:

   $$[[Polska]^{\substack{polski;\ Polska\\ relAdj(placeName.country)}}\ Akademia\ Nauk]^{\substack{Polska\ Akademia\ Nauk\\ orgName}}$$

zostanie zakodowana w następującej postaci:

[B-country@relAdj#B-orgName, I-orgName, I-orgName]

a nazwa *ul. kardynała Stefana Wyszyńskiego* z przykładu (9.9) otrzyma kodowanie:

[B-geogName, I-geogName, I-geogName, B-forename#B-persName#I-geogName, B-surname#I-persName#I-geogName]

Wreszcie dla nazwy nieciągłej z przykładu (9.23) *Gisela Froemel, z domu Kopka, ma 67 lat*, przy wzięciu pod uwagę podtypów i zagnieżdżeń, otrzymujemy kodowanie:

[B-forename#B-persName, B-surname#I-persName, O, O, O, B-surname#I-persName, O, O, O, O]

Metoda ta nosi nazwę Joined Label Tagging (będziemy się do niej odwoływać skrótem JLT).

Metody typu 1 wydają się bardziej obiecujące, lecz w pracy Alex i in. (2007) – w której powyższe techniki zostały szczegółowo opisane – pokazano, że metoda JLT nie ustępuje pozostałym dwóm pod względem osiąganych wyników. Co więcej, JLT wymaga skonstruowania tylko jednego tagera sekwencyjnego, podczas gdy metody klasy 1 wymagają wytrenowania *r* takich tagerów. Do zaanotowania pełnego korpusu jednostkami nazewniczymi wybrana została metoda JLT.

Opiszemy teraz w skrócie algorytm kodowania jednostek w metodzie JLT. Na początek przyjmujemy sekwencję pustych etykiet `O` przypisanych poszczególnym segmentom, `s = [O,O,...,O]`. Kolejne warstwy jednostek kodujemy w kolejności od najbardziej zagnieżdżonej do najmniej zagnieżdżonej (czyli zewnętrznej). Etykiety otrzymane dla kolejnych warstw doklejamy do bieżącego wyniku `s`, przy czym nadmiarowe etykiety `O` są usuwane (tj., etykieta `O#T`, powstała przez połączenie etykiet `T` i `O` z sąsiadujących warstw zagnieżdżenia, zostanie zastąpiona przez `T`). Po przetworzeniu wszystkich warstw wynik kodowania znajduje się na zmiennej `s`.

Proces dekodowania jest odwrotny, należy odkodowywać kolejne warstwy w kolejności od warstwy zewnętrznej do warstwy najbardziej zagnieżdżonej. W trakcie dekodowania może się okazać, że sekwencja etykiet należąca do aktualnie rozważanej warstwy jest niepoprawna, ponieważ tager w ogólnym przypadku nie gwarantuje poprawności sekwencji. W pewnych sytuacjach próbujemy takie sekwencje naprawić. Jeśli etykieta typu `I-T` pojawia się bezpośrednio po etykiecie `O` – i nie było kilka pozycji wcześniej etykiety `B-T`, co by świadczyło o rozpoznaniu nieciągłej jednostki – możemy zamienić ją na etykietę `B-T`. Co zrobić, gdy `I-T` pojawia się w dużej odległości od poprzedzającej ją etykiety `B-T`, a między nimi występują same etykiety puste `O`? Wykorzystywany tager nie modeluje relacji

między odległymi etykietami, więc z dużą dozą pewności możemy stwierdzić, że nie jest to jednostka nieciągła. Przyjmujemy, że odległość między etykietami `B-T` i `I-T` musi być mniejsza niż trzy, aby etykiety zostały potraktowane jak jednostka nieciągła. W przeciwnym wypadku uznajemy, że etykiety te nie są ze sobą związane i reprezentują osobne jednostki.

Jednostki wzajemnie nachodzące

Pozostała jeszcze jedna, potencjalna cecha struktury jednostek nazewniczych: jednostki wzajemnie na siebie nachodzące. Przypomnijmy przyjętą tutaj definicję: mówimy, że jednostki nachodzą na siebie, jeśli segment jednej jednostki wchodzi w obręb drugiej, lub jest położony pomiędzy segmentami drugiej jednostki, jak w przykładach (9.13) i (9.21). Niestety żadnej z tych sytuacji nie da się w prosty sposób odwzorować za pomocą etykiet. W związku z tym, podczas uczenia tagera sekwencyjnego, zachodzące na siebie jednostki są ignorowane, a powstałe narzędzie do rozpoznawania jednostek nazewniczych takich przypadków nie rozpozna. Warto jednak wspomnieć, że nawet gdyby zachodzenie jednostek było na poziomie etykiet reprezentowane, poprawne rozpoznawanie takich przypadków byłoby niezmiernie trudne, ponieważ tagery sekwencyjne z reguły nie radzą sobie dobrze z modelowaniem relacji pomiędzy odległymi od siebie etykietami. W korpusie milionowym około 1% wszystkich jednostek stanowią jednostki wzajemnie na siebie zachodzące.

13.3.2. Tager oparty na CRF

Powyżej opisaliśmy sposób, w jaki zadanie rozpoznawania nazw można sprowadzić do problemu etykietowania sekwencyjnego. Najpierw należy użyć tagera sekwencyjnego do określenia etykiet przypisanych poszczególnym segmentom w zdaniu, a następnie odtworzyć strukturę jednostek nazewniczych z otrzymanej sekwencji etykiet. Sposób kodowania i dekodowania został już wcześniej opisany, teraz opiszemy w skrócie probabilistyczny model *HMM-like linear-chain Conditional Random Field* (CRF; pol. warunkowe pola losowe), wykorzystywany do etykietowania.

Model CRF (Wallach 2004, Sutton i McCallum 2007) określa warunkowy rozkład prawdopodobieństwa $p(y|x)$, gdzie $x = \{x_i\}_{i=1}^{n}$ jest zdaniem, czyli ciągiem segmentów, $y = \{y_i\}_{i=1}^{n}$ jest ciągiem etykiet przypisanych poszczególnym segmentom zdania, a n oznacza długość zdania. Każdy segment x_i reprezentowany jest przez zbiór *obserwacji* określonych dla i-tej pozycji w zdaniu (również na podstawie kontekstu i-tego segmentu). Wybór typów obserwacji jest bardzo

ważny dla jakości etykietowania, a co za tym idzie, również dla jakości rozpoznawania nazw. Przykładowe typy obserwacji, które można brać pod uwagę podczas projektowania narzędzia, wymieniono poniżej:

1. forma ortograficzna i forma podstawowa;
2. prefiksy i sufiksy różnych długości (zarówno form ortograficznych jak i podstawowych);
3. część mowy i kategorie gramatyczne;
4. kształt segmentu – forma ortograficzna przekształcona zgodnie z następującymi regułami:
   a) znaki alfabetu są zamienione na *l* lub *u*, zależnie od wielkości,
   b) cyfry są zastępowane przez *d*,
   c) pozostałe znaki są zamienione na *x*;

   przykładowo, forma *22.02.2011* zostanie zamieniona na *ddxddxdddd*.

Każdy z powyższych typów może dotyczyć nie tylko aktualnego segmentu, ale również segmentów z kontekstu, tak więc obserwacją dla i-tego segmentu w zdaniu może być np. forma podstawowa segmentu na pozycji $i-1$. Zbiór typów obserwacji uwzględnianych podczas konstruowania konkretnego modelu będziemy nazywać *schematem obserwacji*.

Mając zadany schemat obserwacji modelu, należy jeszcze określić wartości parametrów właściwego modelu CRF. Zbiór wartości parametrów będziemy oznaczać przez θ, a warunkowe prawdopodobieństwo przez nie zadane przez $p(y|x,\theta)$. Powstały model, wraz ze schematem obserwacji, może stanowić podstawę do zaanotowania nazwami pełnego korpusu. Wartości parametrów można określić w oparciu o ręcznie zaanotowany jednostkami nazewniczymi korpus milionowy, a dokładniej, w oparciu o *zbiór treningowy* $\mathbf{S} = \{(x^k, y^k)\}_{k=1}^{K}$ skonstruowany na podstawie tego korpusu, gdzie K oznacza liczbę zdań w korpusie. Zbiór treningowy to po prostu zbiór par (zdanie, etykiety), gdzie każde zdanie skonstruowane jest zgodnie z obowiązującym schematem obserwacji – czyli stanowi listę segmentów, z których każdy reprezentowany jest przez zbiór obserwacji – natomiast etykiety stanowią wynik zakodowania jednostek nazewniczych występujących w tym zdaniu. Przykładowo zdanie *Gisela Froemel, z domu Kopka, ma 67 lat* przekształcone zgodnie ze schematem obserwacji `[forma ortograficzna, poprzednia forma ortograficzna, prefiks właściwy długości 3, sufiks właściwy długości 3, kształt]` można przedstawić w sposób pokazany na rys. 13.9.

Każdy z wierszy reprezentuje tu jeden segment przykładowego zdania. Od lewej do prawej zapisane są kolejne obserwacje określone dla poszczególnych segmentów. Na ostatnich pozycjach poszczególnych wierszy znajdują się etykiety powstałe wskutek zakodowania jednostek nazewniczych występujących w zdaniu.

Rysunek 13.9. Reprezentacja zdania *Gisela Froemel, z domu Kopka, ma 67 lat* w zbiorze treningowym modelu CRF

```
W_Gisela PREF_gis SUF_ela SH_llllll              B-forename#B-persName
W_Froemel PW_Gisela PREF_fro SUF_mel SH_lllllll B-surname#I-persName
W_, PW_Froemel SH_x                              O
W_z PW_, SH_l                                    O
W_domu PW_z PREF_dom SUF_omu SH_llll             O
W_Kopka PW_domu PREF_kop SUF_pka SH_llllll       B-surname#I-persName
W_, PW_Kopka SH_x                                O
W_ma PW_, SH_ll                                  O
W_67 PW_ma SH_dd                                 O
W_lat PW_67 SH_lll                               O
```

Dodatkowo, do wartości obserwacji przyłączane są odpowiednie prefiksy: W_, PW_ itd., zależnie od typu obserwacji. Gwarantują one, że wartości różnych typów obserwacji (reprezentowane w postaci napisów) zawsze będą różne i pozwalają identyfikować typ obserwacji na podstawie prefiksu. Oczywiście jest to tylko jeden z wielu możliwych sposobów reprezentacji zdania ze zbioru treningowego.

Wyszukiwanie wartości parametrów modelu będziemy nazywać *trenowaniem modelu*. Celem trenowania jest znalezienie takich wartości parametrów, które maksymalizują dopasowanie modelu do zbioru treningowego $\mathbf{S}$. Bardzo dobre dopasowanie stosunkowo łatwo jest uzyskać, przyjmując schemat o dużej liczbie typów obserwacji, co owocuje modelem CRF o bardzo dużej liczbie parametrów. Jednak taki model niekoniecznie będzie się nadawał do rozpoznawania jednostek nazewniczych w danych, z którymi nie zetknął się podczas trenowania. Zjawisko to nazywamy *przetrenowaniem modelu* (ang. *overfitting*). Aby uniknąć przetrenowania, miara dopasowania, oparta na zlogarytmowanej funkcji wiarygodności, regulowana jest dodatkowym parametrem $1/2\sigma^2$ pozwalającym na kontrolę absolutnych wartości parametrów modelu[12]. Ostateczny wzór na dopasowanie przedstawia się więc następująco:

$$\ell(\theta : \mathbf{S}) = \log\left(\prod_{(x,y)\in\mathbf{S}} p(y|x,\theta)\right) - \sum_{\theta_k\in\theta} \frac{\theta_k^2}{2\sigma^2} \tag{13.10}$$

Tak więc opracowanie tagera sekwencyjnego składa się z dwóch etapów. Najpierw należy okreslić schemat uwzględnianych obserwacji, a następnie wytrenować model w oparciu o powstały zbiór treningowy.

[12] Modyfikacja ta jest równoważna nadaniu parametrom modelu normalnego rozkładu *a priori* $\mathcal{N}(0,\sigma)$. Wytrenowanie modelu polega wtedy na znalezieniu estymatorów maksymalizujących rozkład *a posteriori*.

Projektując schemat obserwacji, ograniczyliśmy kontekst, na podstawie którego określane są wartości obserwacji dla poszczególnych segmentów zdania, do pozycji bieżącej oraz pozycji poprzedniej. Model wytrenowany w oparciu o taki schemat powinien być mniej podatny na przetrenowanie. Ponadto typy obserwacji uwzględniane w modelu zostały ograniczone do tych typów, dla których wartości można uzyskać, korzystając wyłącznie z poziomu segmentacji i poziomu tekstu. Korpus milionowy został ręcznie poprawiony, dlatego poziomy morfoskładniowe w korpusie milionowym i w korpusie pełnym charakteryzują się różną jakością, informacje pochodzące z poziomu morfoskładniowego – np. części mowy – nie są więc w schemacie brane pod uwagę.

Schemat obserwacji wykorzystany do zbudowania finalnej wersji tagera wygląda następująco:

1. forma ortograficzna, w której duże litery zamienione są na małe; forma z zachowanymi dużymi literami (jeśli występują) stanowi osobną obserwację;
2. dla każdej formy ortograficznej, jej prefiksy i sufiksy długości $k, k-1, k-2, k-3$ (o ile niepuste), gdzie k to długość formy; obserwacje te stanowią substytut form podstawowych; wszystkie duże litery zamienione są na małe;
3. kształt formy ortograficznej;
4. skompresowany kształt formy ortograficznej; sąsiadujące, identyczne znaki zostają zamienione na jeden znak; przykładowo, kształt *ulllxx* po skompresowaniu będzie miał wartość *ulx*;
5. dla każdej formy ortograficznej, jej sufiksy właściwe długości $3, 4, 5$; obserwacje te mogą pomóc w rozpoznawaniu derywacji, szczególnie przymiotników relacyjnych, które charakteryzują się podobnymi końcówkami;
6. połączone kształty form ortograficznych sąsiadujących segmentów;
7. połączone skompresowane kształty form ortograficznych sąsiadujących segmentów.

Obserwacje opisane w trzech ostatnich podpunktach dotyczą słowa bieżącego. Pozostałe obserwacje każdorazowo dotyczą słowa bieżącego i poprzedniego.

13.3.3. Wyniki

Wyniki rozpoznawania nazw opiszemy w dwóch częściach. Najpierw skupimy się na korpusie milionowym, na którym została przeprowadzona ewaluacja narzędzia. W drugiej części podamy liczbę jednostek nazewniczych znalezionych przez narzędzie w pełnym korpusie.

W celu dokonania oceny jakości metody rozpoznawania jednostek, na korpusie milionowym (w wersji z 20 czerwca 2011 r.) składającym się z 85 425 zdań, 1 211 320 segmentów oraz 86 599 jednostek nazewniczych została przeprowadzona *walidacja krzyżowa* (ang. *cross-validation*). Korpus został podzielony na pięć części w taki sposób, aby do każdej części trafiła mniej więcej taka sama liczba jednostek nazewniczych (około 17 300). Liczba zdań w poszczególnych częściach wahała się od 16 500 do 17 400. Co ważne, podział został przeprowadzony na poziomie plików, nie zdań – zdania z jednego pliku korpusu milionowego zawsze trafiały do tej samej części. Dla każdej z uzyskanych części o numerze $k \in \{1, \ldots, 5\}$:

1. Na zbiorze treningowym skonstruowanym w oparciu o pozostałe cztery części $\{1, \ldots, 5\} \setminus \{k\}$ wytrenowany został model CRF.
2. Na części o numerze k obliczone zostały statystyki oceniające jakość rozpoznawania nazw przy użyciu powstałego modelu – dokładność, pełność oraz miara F, czyli średnia harmoniczna dokładności i pełności.

Wynikiem walidacji krzyżowej są uśrednione wartości podanych wyżej statystyk obliczone względem podziału korpusu milionowego. Są one przedstawione w tab. 13.2. Podobnie jak przy wyliczaniu wyników SProUT-a oraz zgodności anotatorów, poszczególne typy danych połączone zostały w kategorie ułatwiające porównanie z innymi pracami. Tak więc *nazwy osobowe* odpowiadają typowi `persName`, *nazwy miejscowe* obejmują typy `geogName` i `placeName`, *nazwy organizacji* odnoszą się do typu `orgName`, *wyrażenia czasowe* do `date` i `time`, wreszcie *derywacje* oznaczają jednostki mające cechy `derivType` i pochodzące od nazw dowolnego typu i podtypu.

Tabela 13.2. Wyniki walidacji krzyżowej względem podziału na 5 części. Należy zwrócić uwagę, że ostatnia kolumna podaje średnią wyników dla miary F, a nie średnią harmoniczną dwóch poprzednich kolumn

Typ	**Uśrednione statystyki**		
	Dokładność	Pełność	Miara F
Nazwy osobowe	0,86	0,80	0,83
Nazwy miejscowe	0,83	0,71	0,77
Nazwy organizacji	0,70	0,65	0,67
Wyrażenia czasowe	0,86	0,80	0,83
Derywacje	0,87	0,68	0,77
Ogółem	0,83	0,76	0,79

Tabela 13.2 pokazuje, że – podobnie jak było to w wypadku SProUT-a (p. 13.1.4) – najwyższe wyniki otrzymywane są tu dla wyrażeń czasowych, a najniższe dla nazw organizacji. Dodatkowo w wypadku CRF wyniki dla nazw osobowych są równie dobre jak dla wyrażeń czasowych. Dla wszystkich kategorii

dokładność jest wyższa od pełności. Wyniki dla CRF są zdecydowanie lepsze od otrzymanych dla SProUT-a. Dzięki danym zawartym w pracy Savary i Piskorskiego (2011), możliwe jest szczegółowe porównanie tych dwóch narzędzi w sytuacji, gdy brane są pod uwagę wszystkie jednostki, w tym zagnieżdżone, ich typy, podtypy oraz typy derywacji, lecz pomijane są ich formy podstawowe i znormalizowane oraz bazy derywacyjne (por. p. 9.3.2). Różnice wahają się, zawsze na korzyść CRF: w wypadku dokładności od 3% dla wyrażeń czasowych do 8% dla nazw organizacji, a w wypadku pełności od 11% dla derywacji do 46% dla nazw osób. Pamiętajmy jednak, że ten spektakularny dla uczenia maszynowego bilans analizy porównawczej możliwy jest dzięki dużemu nakładowi pracy ręcznej anotatorów (ponad 10 osobomiesięcy), mimo wszystko istotnie zredukowanemu właśnie dzięki użyciu SProUT-a do anotacji wstępnej. Istotną przewagą SProUT-a jest również zdolność do proponowania form podstawowych i baz derywacyjnych dla oznaczanych jednostek, czego nie potrafi obecny model CRF.

Tabela 13.3. Liczba oznaczonych jednostek nazewniczych w pełnym korpusie

Nazwy osobowe	**Nazwy organizacji**	**Nazwy geograficzne**	**Nazwy geopolityczne**
94 991 096	22 593 467	7 736 442	28 666 309
Daty	**Godziny**	**Przymiotniki relacyjne**	**Derywacje osobowe**
9 778 340	2 252 954	9 375 486	1 834 876
Wszystkie jednostki nazewnicze			166 018 608

Tabela 13.3 pokazuje liczbę nazw poszczególnych kategorii oznaczonych za pomocą modelu CRF w pełnym korpusie. Nie dysponujemy oczywiście danymi co do jakości tej anotacji. Można zauważyć, że – podobnie jak dla korpusu milionowego (por. tab. 9.1) – najliczniejsze są nazwy osób. Przypomnijmy jednak, że specyficzne reguły anotacyjne dotyczące ich stopnia zgnieżdżania niekiedy sztucznie zawyżają ich liczbę. Dwie kolejne najliczniejsze kategorie to nazwy geopolityczne i nazwy organizacji. Najrzadziej zaś występującymi typami są derywacje osobowe i określenia godzin.

13.3.4. Analiza błędów

Wyniki walidacji krzyżowej przedstawione w tab. 13.2 zawierają podstawowe statystyki rozpoznawania jednostek w podziale na główne klasy jednostek nazewniczych. W celu przeprowadzenia dokładniejszej analizy błędów korpus milionowy został podzielony (na poziomie plików, tak jak w wypadku walidacji krzyżowej) na część treningową oraz część testową, przy czym część treningowa stanowiła 90% całego korpusu. Na części treningowej wytrenowany został model

CRF, natomiast analiza dotyczy błędów popełnianych w trakcie rozpoznawania jednostek w części testowej przy użyciu uzyskanego modelu. Analizę przeprowadzimy w oparciu o listy klas najczęściej popełnianych błędów. Określenie i umiejętne usunięcie przyczyn występowania najczęstszych błędów prowadziłoby do największego wzrostu jakości rozpoznawania jednostek nazewniczych.

Na początek rozważmy fałszywe pozytywy oraz fałszywe negatywy, dla których statystyki przedstawione są w tab. 13.4. Na potrzeby analizy błędów pojęcia te będziemy rozumieli w następujący sposób: fałszywe pozytywy to jednostki rozpoznane w miejscach, w których nie ma w korpusie milionowym żadnej jednostki (również jednostki o innym typie lub częściowo pokrywającym się zbiorze segmentów); fałszywe negatywy to zjawisko odwrotne – zachodzi ono, gdy w miejscu występującej w korpusie jednostki nie została rozpoznana żadna jednostka. Statystyki nie są zaskakujące: najwięcej fałszywych pozytywów i negatywów występuje wśród nazw osobowych, które pojawiają się w korpusie najczęściej, oraz nazw organizacji, które również są liczne, a jednocześnie najgorzej rozpoznawane przez narzędzie (por. tab. 13.2). Grupa fałszywych negatywów ma dość jednolity charakter – narzędzie nie rozpoznało jednostek, z którymi nie spotkało się w korpusie treningowym, a kontekst i obserwacje dotyczące wystąpienia jednostki w korpusie testowym również nie są wystarczające do jej rozpoznania. Fałszywe pozytywy natomiast to zwykle jednostki typów nieanotowanych w korpusie NKJP, jak we frazach *publikuje słynną książkę „Our Bodies Our Selves"* – tytuł książki, *jeśli WebScan znajdzie wirusa* – nazwa oprogramowania, *taki komunikat ogłosiło Walne Zgromadzenie Wspólników* – nazwa wydarzenia, *grób Pikusia i Miśka rozświetlają znicze* – tu: imiona psów, itp. Fałszywe pozytywy pojawiają się również na początku zdania – na przykład we frazie *Lot samolotem traktuję*, gdzie słowo *Lot* zostało uznane przez narzędzie za jednostkę nazewniczą.

Tabela 13.4. Najczęściej występujące fałszywe negatywy oraz fałszywe pozytywy

Oczekiwany typ lub podtyp	**Fałszywe negatywy**
Nazwa osobowa (bez podtypu)	307
Nazwa organizacji	202
Nazwisko	186
Imię	121
Nazwa geograficzna	85

Rozpoznany typ lub podtyp	**Fałszywe pozytywy**
Nazwa organizacji	95
Nazwa osobowa (bez podtypu)	65
Nazwisko	46
Imię	43
Nazwa miejscowości	21

Ciekawsze są przypadki błędnego określania typów, podtypów lub pokrycia segmentów rozpoznanych jednostek. Jak widać w tab. 13.5, najczęściej popełnianym błędem tego rodzaju jest niepoprawne określenie zakresu rozpoznanej nazwy organizacji. We frazach *Amerykańska Komisja Równych Szans*, *Bostoński*

Kolektyw Zdrowia Kobiet i *Wielkopolskich Zakładów Przemysłu Tłuszczowego w Szamotułach* narzędzie rozpoznało pierwsze słowa – *Amerykańska*, *Bostoński* i *Wielkopolskich* – jako osobne, niewchodzące w skład całej nazwy organizacji jednostki. Tutaj błędne określenie zakresu związane jest ze złożoną strukturą rozpoznawanej jednostki – niewłączone w skład nazwy organizacji słowa stanowią zagnieżdżone przymiotniki relacyjne. Podobnie dla frazy *Uniwersytetu Jagiellońskiego w Krakowie* jedynie *Uniwersytet Jagielloński* został rozpoznany przez narzędzie jako nazwa organizacji. Czasami występuje również sytuacja odwrotna, gdy zakres oznaczonej jednostki jest zbyt rozległy. Przykładowo cała fraza *Ursuska „S"* została rozpoznana jako nazwa organizacji, podczas gdy w korpusie milionowym jedynie *„S"* stanowi nazwę organizacji, natomiast *Ursuska* jest osobnym przymiotnikiem relacyjnym. Inną sytuację obrazuje przykładowa fraza *Brytyjski szef MSZ Robin Cook*, w której sąsiadujące słowa *MSZ Robin Cook* narzędzie uznało za pojedynczą nazwę organizacji. Warto zauważyć, że dyskryminujące byłyby tutaj wartości przypadków kolejnych słów, które prawdodpodobnie pozwoliłby poprawnie rozpoznać jednostki występujące we frazie. Obecny model nie korzysta jednak z informacji z poziomu morfoskładniowego. Gdyby w podanej frazie zamiast słowa *Cook* pojawiło się bardziej znane nazwisko, np. *Smith*, narzędzie również rozpoznałoby poprawnie wszystkie występujące we frazie jednostki. Innym rodzajem błędu jest łączenie sąsiadujących nazw organizacji, rozdzielonych spójnikiem *i* – który często stanowi część składową nazwy organizacji – jak we frazie *Klubem Paryskim i Klubem Londyńskim*.

Tabela 13.5. Najczęściej popełniane przez narzędzie błędy w korpusie testowym, ze względu na oczekiwany i rozpoznany typ jednostki. Jeśli typy się zgadzają, występuje błąd określenia zakresu jednostki

Oczekiwany typ lub podtyp	**Rozpoznany typ lub podtyp**	**Liczba błędów**
Nazwa organizacji	nazwa organizacji	131
Nazwa organizacji	nazwa państwa	55
Imię	nazwisko	47
Nazwa organizacji	nazwa miejscowości	46
Nazwa osobowa (bez podtypu)	nazwa osobowa (bez podtypu)	42
Data	data	35
Nazwisko	imię	34

Drugą najliczniejszą klasą błędów jest rozpoznawanie nazwy państwa lub kraju w miejscu, gdzie występuje nazwa organizacji. Jest to związane z metonimią opisaną szczegółowo w p. 9.4.1. Wśród tego typu błędów dominują słowa takie

jak *ZSSR*, *Rosji*, *Polski*, *Izrael*, *Ukrainą* itp. W większości tych przypadków kontekst semantyczny byłby niezbędny do określenia poprawnej interpretacji danego słowa. Innym ciekawym przypadkiem jest słowo *Polamu* rozpoznane jako nazwa państwa, zapewne na zasadzie podobieństwa do nazwy *Polska*. Przypomnijmy, że podobieństwo ortograficzne wyraża się m.in. przez wartości prefiksów (które wchodzą w skład finalnego schematu obserwacji).

Kolejną liczną grupą błędów są wzajemnie mylone imiona i nazwiska, szczególnie gdy występują w zdaniu pojedynczo. Przykładami fraz, dla których podtyp jednoczłonowej nazwy osobowej nie został poprawnie określony, są: *Prof. Macieja przypomniał, to ostry głos Dragana*, *przytaknął Marciniak* lub *dość mam homilii Xaviera*. Do trudniejszych przypadków, z którymi model sobie nie radzi, należą również trójczłonowe nazwy osobowe składające się z dwóch imion i nazwiska, takie jak *Lin Jing Hua* albo *Lecha Emfazego Stefańskiego*. Często występującym w korpusie przypadkiem są dwuczłonowe nazwiska, stąd gdy drugie imię – *Jing*, *Emfazy* – nie jest zbyt częste w korpusie treningowym (lub pojawia się równolegle jako nazwisko), może zostać rozpoznane przez narzędzie jako nazwisko.

Do czwartej z najliczniejszych klas należą rozpoznawane, zamiast organizacji, nazwy miejscowości. Podobnie jak w drugiej z opisywanych klas, tutaj również często trudno odróżnić – bez dodatkowej, semantycznej informacji – miejscowość od organizacji. Przykładami mogą być frazy: *czy Kalisz znajdzie miejsce na zbiory*, *Czy za rok w Mielcu będzie kolejny strajk* lub *dla Pekinu dysydent za granicą jest lepszy niż dysydent w więzieniu*. W odróżnieniu od nazw państw, nazwy miejscowości dużo częściej pojawiają się (i stanowią dominujący fragment) w nazwach organizacji, co jest kolejnym źródłem błędów należących do tej klasy. We frazach *znalazłeś się w Djurgaarden Sztokholm*, *skrzydłowy New Jersey Nets* i *sympatycy Toronto Raptors* żadna organizacja nie została rozpoznana przez narzędzie poprawnie.

Na kolejną grupę składają się błędy dotyczące źle określonego zakresu nazwy osobowej. Błędy te zwykle wynikają z nierozpoznania imienia, nazwiska lub przydomku, szczególnie w przypadkach dwuczłonowych nazwisk lub nazw osobowych z podanymi dwoma (lub więcej) imionami. Przykładowo, w nazwie *Bożena Nowak - Szymura* jedynie *Bożena Nowak* została rozpoznana jako nazwa osobowa, a w nazwie *Timothy'ego Gartona Asha* słowo *Asha* również nie zostało włączone w skład całej nazwy osobowej. We frazie *Wiaczesława Iwankowa alias Japończyka* słowo *Japończyka* nie zostało rozpoznane jako przydomek – zamiast tego, zostało oznaczone jako nazwa mieszkańca Japonii – i w związku z tym nie zostało włączone w skład nazwy osobowej.

Do ostatniej z rozpatrywanych grup należą błędy dotyczące określenia zakresu daty. Dla przykładu, we frazie *kwietniu 1981 r.* słowo *kwietniu* nie zostało włączone w obręb jednostki nazewniczej, natomiast we frazie *osiemdziesiątego*

piątego roku narzędzie rozpoznało datę w postaci *osiemdziesiątego piątego*, bez słowa *roku*. Ponadto w wypadku dat często występuje wzajemne nachodzenie na siebie jednostek (jak wspomnieliśmy, narzędzie nie modeluje tego zjawiska). We frazie *w latach 2025–30* rozpoznana przez narzędzie została tylko pierwsza data w postaci *latach 2025*, natomiast pominięta została druga z nich, *latach 30*. Podobna sytuacja ma miejsce we frazach *7 do 11 marca* oraz *1866 i 1870 roku*.

13.4. Wnioski i perspektywy

Przedstawiliśmy w tym rozdziale najważniejsze narzędzia używane podczas anotacji jednostek nazewniczych w korpusie NKJP. W celu usprawnienia anotacji ręcznej podkorpusu milionowego dokonywaliśmy automatycznie jego anotacji wstępnej za pomocą platformy SProUT opartej na zasobach lingwistycznych. Istniejące wcześniej słowniki dziedzinowe i gramatyki do ekstrakcji polskich jednostek nazewniczych zostały przez nas rozszerzone i dopasowane do zadania anotacji. Tak otrzymane narzędzie pozwala na uzyskanie dokładności wahającej się w granicach od 68% do 78% oraz pełności od 35% do 39%.

Ręczna poprawa i uzupełnianie anotacji wstępnej dokonywane były za pomocą edytora drzew TrEd, dostosowanego do naszych potrzeb przez rozszerzenia, skróty klawiaturowe i arkusze stylów. Każdy plik podkorpusu milionowego znakowany był w ten sposób przez dwóch niezależnych anotatorów, a niezgodności ich decyzji rozstrzygane były przez superanotatora. Proces ten trwał przez 10 miesięcy – od kwietnia 2010 do stycznia 2011 r., a ogólna zgodność anotatorów wynosiła 83%.

Podkorpus milionowy stał się następnie podstawą do stworzenia prototypu narzędzia opartego o metody uczenia maszynowego. W trakcie trenowania modelu CRF brane były pod uwagę cechy każdego segmentu podkorpusu milionowego i segmentu go poprzedzającego, takie jak formy ortograficzne, prefiksy, sufiksy i kształty słów. W identyfikacji jednostek nazewniczych oraz określaniu ich typów, podtypów i typów derywacji narzędzie to osiągnęło wyniki rzędu 83% dokładności i 76% pełności. Są to wyniki istotnie lepsze niż uzyskane przez SProUT-a. Dlatego też do anotacji pełnego korpusu zdecydowano się użyć właśnie tego prototypu, co zaowocowało oznaczeniem ponad 166 milionów jednostek nazewniczych. Zestaw ich cech nie uwzględnia jednak form podstawowych i baz derywacyjnych, dlatego interesującą perspektywą byłoby stworzenie narzędzia hybrydowego, a więc opartego zarówno na wiedzy jak i na uczeniu maszynowym. Na przykład istniejący moduł CRF pozwalałby na identyfikację i kategoryzację jednostek, a gramatyki i słowniki dziedzinowe umożliwiałyby określanie ich form podstawowych i baz derywacyjnych.

14

Wyszukiwarka PELCRA dla danych NKJP

Piotr Pęzik

14.1. O wyszukiwarce

Wyszukiwarka PELCRA[1] to jedna z dwóch wyszukiwarek internetowych dostępnych dla danych NKJP. Narzędzie to pozwala na szybkie i wygodne przeszukiwanie zasobów zgromadzonych na potrzeby Narodowego Korpusu Języka Polskiego. Wyszukiwarka jest oparta na składni zapytań korpusowych, która z jednej strony oferuje funkcjonalność porównywalną z opcjami wyszukiwania dostępnymi w standardowych narzędziach korpusowych, a z drugiej umożliwia szczególnie skuteczne wyszukiwanie pojedynczych wyrazów, wariantów morfologicznych i semantycznych oraz elastycznych wielowyrazowych kolokacji w obszernych zbiorach danych NKJP. W czasie powstawania niniejszej publikacji wyszukiwarka pozwalała na bardzo szybkie przeszukiwanie ponad 1500 milionów słów tekstowych znajdujących się w ogólnej puli danych NKJP. Dla przykładu, niezapisane w pamięci podręcznej serwera zapytanie o wszystkie wystąpienia różnych form rzeczownika *brama*, występującego w ponad 54 tysiącach kontekstów korpusowych zajmuje niecałe 0,2 sekundy. Co ważne, wyszukiwarka PELCRA zawsze podaje całkowitą liczbę kontekstów wystąpień wyrazów, lub fraz pasujących do danego zapytania, nawet dla bardzo częstych wyrazów lub fraz.

[1] PELCRA to akronim nazwy zespołu badawczego działającego od 1997 roku w Instytucie Anglistyki Uniwersytetu Łódzkiego – zob. `http://pelcra.ia.uni.lodz.pl`. Opisana w tym rozdziale wyszukiwarka dla danych NKJP, zwana po prostu *wyszukiwarką PELCRA*, dostępna jest pod adresem `http://nkjp.uni.lodz.pl/`.

Inną wyszukiwarką dostępną od wczesnych etapów projektu jest internetowa wersja silnika Poliqarp (Janus i Przepiórkowski 2007)[2], która została przystosowana do przeszukiwania danych NKJP. Do niewątpliwych zalet tej wyszukiwarki należy jej bogata składnia, pozwalająca formułować między innymi zaawansowane zapytania uwzględniające kryteria morfosyntaktyczne, które wykorzystują anotację lingwistyczną całego korpusu. Szczegółowy opis funkcji Poliqarpa dla NKJP zostały opisane na stronach pomocy tej wyszukiwarki (`http://nkjp.pl/poliqarp/help/pl.html`). Składnia Poliqarpa została szczegółowo opisana w Przepiórkowski 2004.

14.2. Skrócone odsyłacze

Zanim przejdziemy do omawiania poszczególnych funkcji wyszukiwarki PELCRA, warto wprowadzić opcję generowania skompresowanych odsyłaczy do wyników. Aby ułatwić użytkownikom odtwarzanie wyników wysyłanych do wyszukiwarki zapytań można za pomocą przycisku `URL` stworzyć krótki odsyłacz do bieżącego ekranu zapytania. Odsyłacz zostaje wyświetlony tuż pod oknem wyszukiwania (zob. rys. 14.1).

Rysunek 14.1. Przykład skróconego odsyłacza

W skompresowanym odsyłaczu zakodowane są wszystkie informacje o wybranych opcjach wyszukiwania. Po jego kliknięciu wyświetlany zostanie nie tylko ekran zapytania, ale też wyniki, które zwraca dane zapytanie. Taki skompresowany odsyłacz można łatwo zapisać, zamieścić w publikacji, lub przesłać pocztą elektroniczną. Przy większości przykładów omawianych poniżej podano bezpośredni skrócony odsyłacz do ekranu wyników pasujących do danego zapytania, dzięki czemu czytelnik może sprawdzić działanie zapytań.

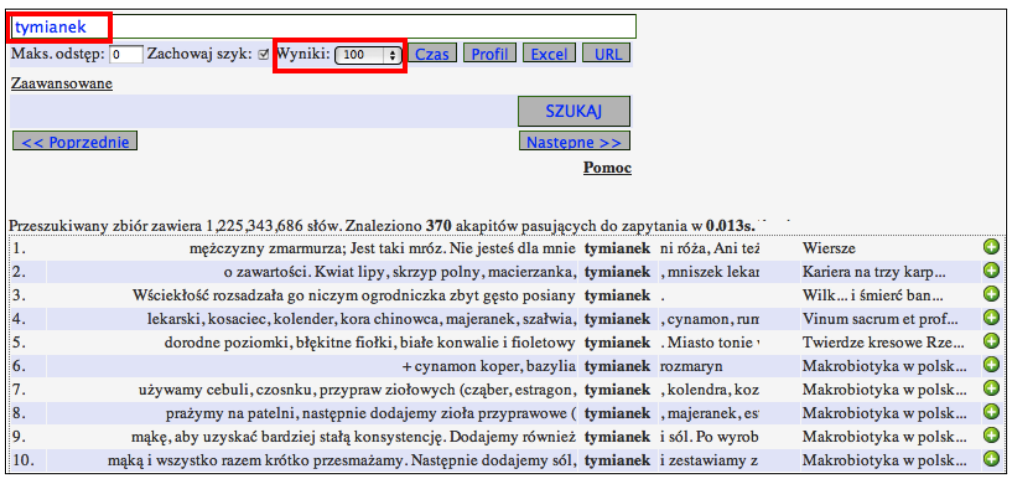
[2] W ramach NKJP Poliqarp był rozwijany i utrzymywany przez Jakuba Wilka.

14.3. Składnia zapytań w przykładach

14.3.1. Wyszukiwanie dokładnych dopasowań pojedynczych wyrazów

Aby wyszukać wystąpienia danego słowa w korpusie, należy je wpisać w szerokim polu tekstowym na górze formularza zapytania. Po kliknięciu przycisku `Szukaj` wyświetlą się wystąpienia tego słowa w zindeksowanym korpusie. Na przykład po wpisaniu wyrazu *tymianek*[3] powinny się ukazać konteksty zawierające jego dokładne dopasowania (rys. 14.2)[4].

Rysunek 14.2. Przykład dokładnego dopasowania dla wyrazu *tymianek*



Ogólna liczba kontekstów pasujących do zapytania jest podawana bezpośrednio nad tabelą wyników. Wyniki można *posortować* według dopasowania (ma to sens w przypadkach opisywanych poniżej), lub też według *lewego* albo *prawego* słowa w konkordancji. Możliwe jest też określenie *maksymalnej liczby wyników* pojawiających się na stronie. Przechodzenie między kolejnymi stronami wyników umożliwiają przyciski << *Poprzednie* oraz *Następne* >>. Informacje o tekście, z którego pochodzi dany cytat, a także szerszy kontekst wystąpienia można uzyskać poprzez kliknięcie symbolu zielonego kółka z plusem w ostatniej kolumnie danego wiersza wyników.

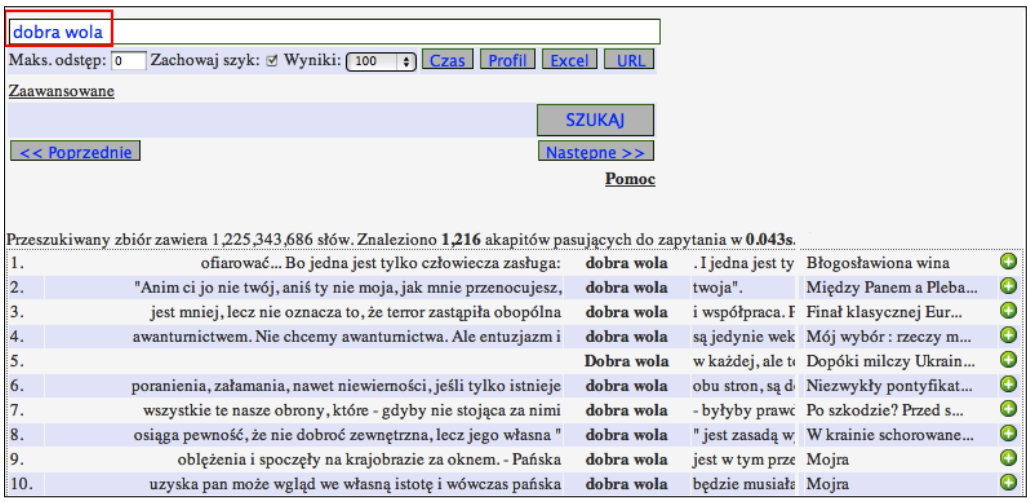
[3] Zob. `http://nkjp.uni.lodz.pl/?q=yhfmpxr`.

[4] `Dopasowaniem` nazywamy tu wystąpienie słowa lub frazy w korpusie, które pasuje do zapytania użytkownika.

14.3.2. Wyszukiwanie dokładnych dopasowań fraz

Aby wyszukać frazę dokładnie pasującą do zapytania, należy ją wpisać w pole zapytania, zaznaczyć opcję *Zachowaj szyk* oraz określić maksymalny *odstęp*[5] między wyrazami wartością 0 (rys. 14.3).

Rysunek 14.3. Przykład dokładnego dopasowania frazy *dobra wola*



14.3.3. Ortograficzne symbole wieloznaczne

Składnia obsługuje kilka rodzajów symboli wieloznacznych. Dwa podstawowe symbole wieloznaczne, tj. * (0 lub więcej dowolnych znaków) oraz ? (jeden dowolny znak) umożliwiają ortograficzne rozszerzenie terminu zapytania. Na przykład zapytanie `tymian*`[6] zwraca dopasowania *tymianek*, *tymiankowy*, ale też *Tymiankach*. Z kolei zapytanie `osobliw?` zwraca dopasowania wyrazów *osobliwy*, *osobliwa*, *osobliwą* itd.

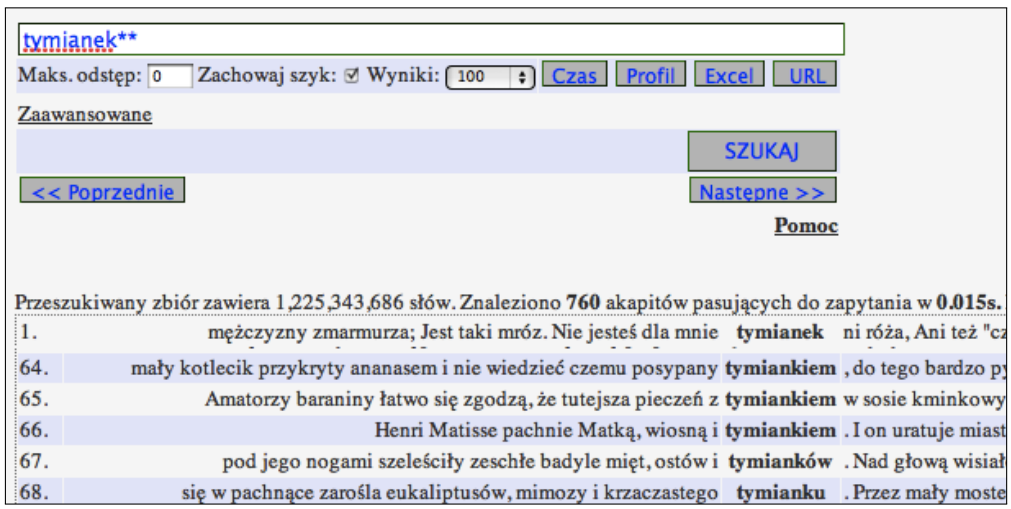
14.3.4. Wyszukiwanie fleksyjne słownikowe

W językach bogatych fleksyjnie ortograficznie rozszerzone zapytanie może zwracać mało dokładne konkordancje, w których warianty fleksyjne są przemieszane z derywatami należącymi do innej kategorii części mowy.

[5] Zob. `http://nkjp.uni.lodz.pl/?q=yhvcekc`.
[6] Zob. `http://nkjp.uni.lodz.pl/?q=62t4z2f`.

Dlatego w wyszukiwarkach tworzonych dla korpusów polszczyzny bardzo przydatna jest możliwość wyszukiwania fleksyjnego. Opisywana tu wyszukiwarka obsługuje prosty, ale bardzo przydatny rodzaj wyszukiwania fleksyjnego, tzn. wyszukiwanie fleksyjne słownikowe (na podstawie słownika Morfologik, `http://morfologik.blogspot.com/`). Aby automatycznie rozszerzyć zapytanie o warianty fleksyjne zadanej formy podstawowej (np. rzeczownika w mianowniku, rodzaju męskim w liczbie pojedynczej), należy na końcu takiej formy dodać symbol podwójnej gwiazdki (`**`). Na przykład zapytanie `tymianek**` może zwrócić zbiór dopasowań ukazany na rys. 14.4.

Rysunek 14.4. Przykład wyszukiwania fleksyjnego słownikowego dla zapytania `tymianek**`

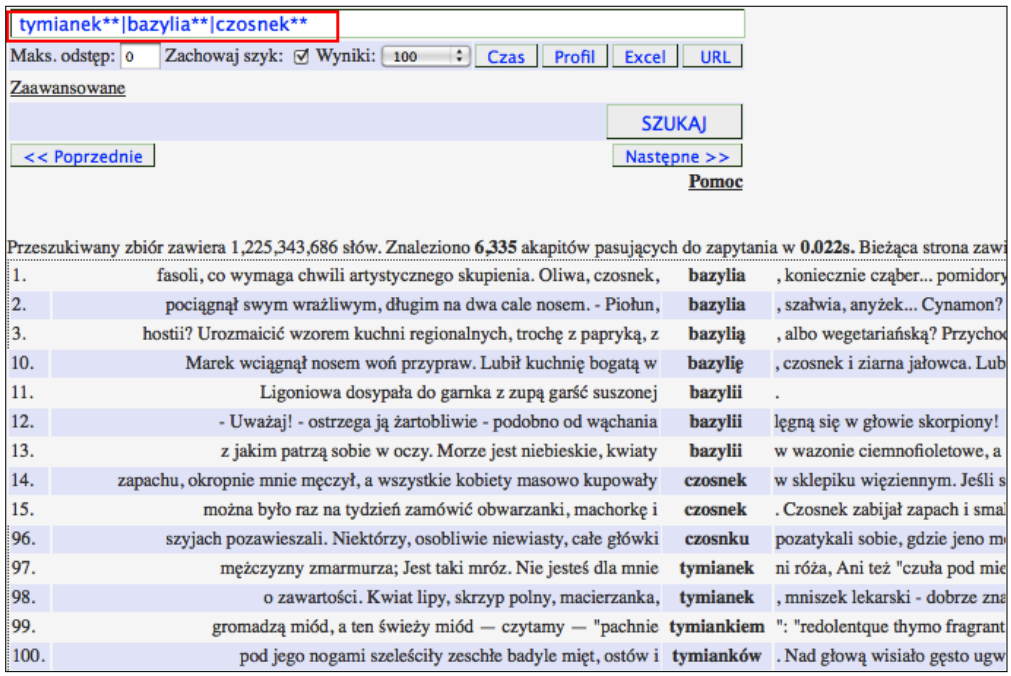


14.3.5. Wyszukiwanie wariantów

Składnia wyszukiwarki umożliwia również formułowanie zapytań zawierających warianty morfologiczne, zbiory synonimów lub nawet antonimy określane przez autora zapytania. Użycie symbolu `|` spowoduje, że dopasowane zostaną wystąpienia dowolnego z wyrazów w danej grupie wariantów. Na przykład zapytanie `tymianek**|bazylia**|czosnek**`[7] zwróci wystąpienia dowolnego z tych trzech wyrazów. W tym wypadku dopasowane zostaną również ich odmiany, ze względu na użyty symbol podwójnej gwiazdki (rys. 14.5).

[7] Zob. `http://nkjp.uni.lodz.pl/?q=p9knmh`.

Rysunek 14.5. Przykład wyszukiwania wariantów morfologicznych dla zapytania `tymianek**|bazylia**|czosnek*`



14.3.6. Rozszerzenie ortograficzne na początku wyrazu

Wyszukiwarka obsługuje również zapytania z „gwiazdką" na początku wyrazu. Na przykład zapytanie `*filetow*` zwróci wystąpienia wyrazów *sfiletować*, *odfiletować* oraz *wyfiletować*. Z kolei zapytanie `*essa**` zwróci wszystkie odmiany występujących w słowniku morfologicznym wyrazów zakończonych przyrostkiem *-essa*, czyli na przykład *stewardessa*, *poetessa*, *hostessa*.

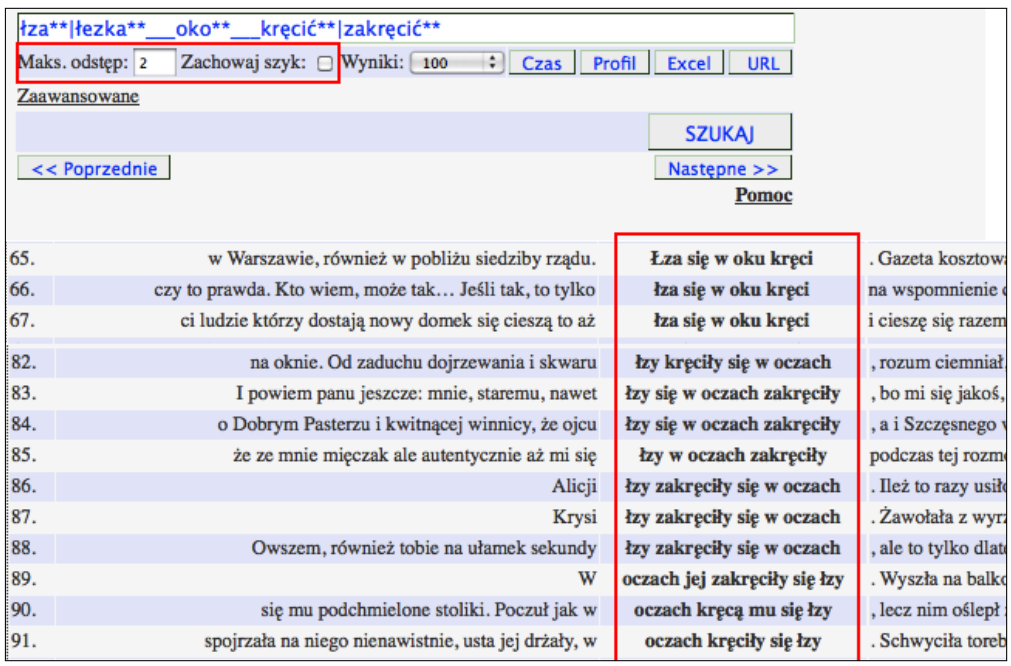
14.3.7. Dopasowywanie elastycznych związków frazeologicznych poprzez wyszukiwanie kontekstowe

Składnia wyszukiwarki umożliwia szczególnie wygodne wyszukiwanie wielowyrazowych związków frazeologicznych, które często cechują się luźnym szykiem wyrazów. Aby wyszukać kolokacje rzeczowników *łza* oraz/lub *łezka* z rzeczownikami *oko* w żądanym kontekście we wszystkich odmianach tych wyrazów, należy sformułować zapytanie `łza**|łezka**___oko**` (grupy wariantów są tu rozdzielone potrójnym podkreślnikiem). Maksymalny odstęp między terminami zapytania możemy dla przykładu określić wartością 2, przy czym zaznaczenie

opcji *Zachowaj szyk* ograniczy liczbę dopasowań do kontekstów, w których wyrazy występują w kolejności ich podania w zapytaniu.

Podobne, choć nieco bardziej uściślone zapytanie `łza**|łezka**___oko**_-__kręcić**| zakręcić**`[8] może zwrócić zbiór wyników ukazany na rys. 14.6.

Rysunek 14.6. Przykład dopasowywania elastycznych związków frazeologicznych poprzez wyszukiwanie kontekstowe



Warto zwrócić uwagę na to, że w niektórych wyszukiwarkach korpusowych dopasowanie tak elastycznego związku frazeologicznego wymagałoby sformułowania co najmniej kilku osobnych zapytań dla poszczególnych wariantów.

14.4. Sortowanie

Opcje sortowania oraz grupowania wyników dostępne są w zaawansowanym formularzu wyszukiwania. Zbiory wyników można sortować dwustopniowo (np. najpierw według źródła, a następnie daty) według następujących kryteriów:

1. Dopasowanie (środek). Sortowanie według dopasowania ułatwia analizę konkordancji wariantów ortograficznych i morfologicznych. Na przykład posortowanie wyników zapytania `ręka**` według dopasowania podzieli

[8] Zob. `http://nkjp.uni.lodz.pl/?q=p33wg5`.

konkordancje na podzbiory zawierające wystąpienia różnych form rzeczownika *ręka*.
2. Lewy lub prawy kontekst. Sortowanie według kontekstu umożliwia prostą analizę najczęstszych kolokacji pozycyjnych występujących w zbiorze wyników.
3. Źródło. Źródłem w wypadku tekstów gazetowych jest tytuł gazety (ale nie pojedynczego artykułu), a w wypadku książek ich tytuł.
4. Data publikacji.
5. Kanał, np. prasa, książka, Internet, nagrania języka mówionego.

Warto podkreślić, że sortowane są tylko zbiory wyników (maks. 10 000 na raz), a nie wszystkie wystąpienia w korpusie.

14.5. Grupowanie

Pewnych problemów przy analizowaniu wyników konkordancji z dużych korpusów nastręczają powtórzenia wystąpień częstych wyrazów w tych samych gazetach, książkach lub też w tekstach z tego samego okresu. Często użytkownika korpusu interesują przykłady użycia danego wyrazu lub frazy w różnych gazetach, tekstach, latach, podczas gdy zbiory niepogrupowanych wyników mogą zawierać nadmiar przykładów z jednego źródła.

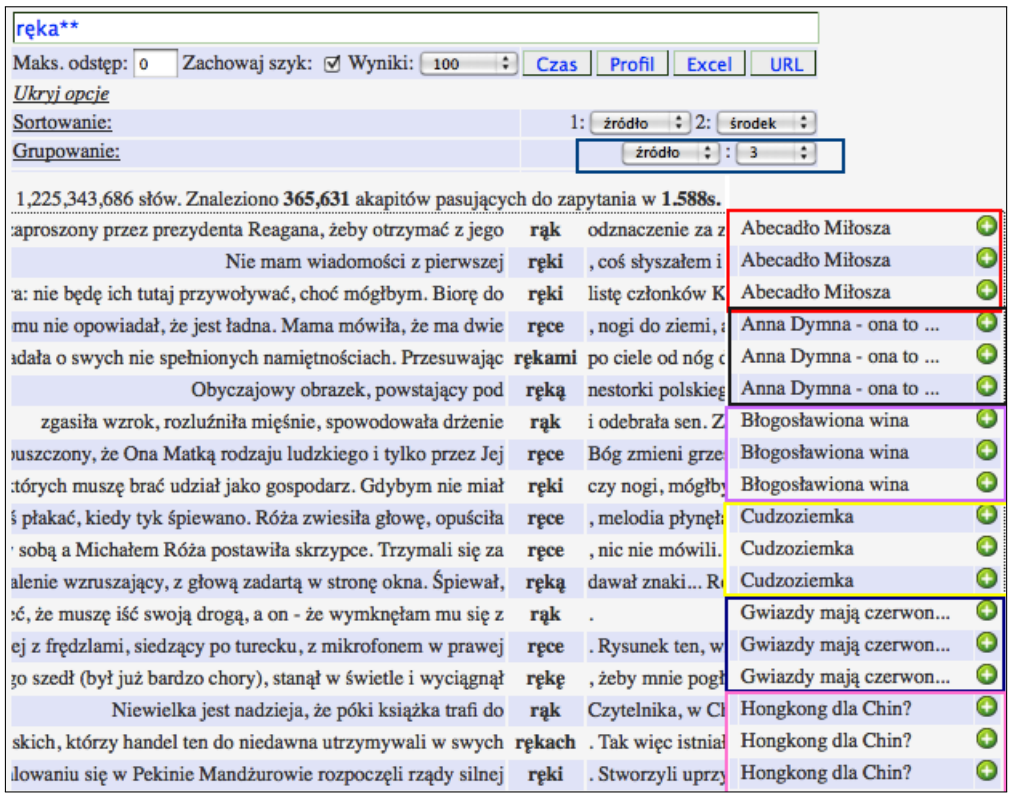
Opcja grupowania wyników umożliwia określenie maksymalnej liczby wyników z danego roku, źródła lub tekstu. Widać to na ukazanym na rys. 14.7 przykładzie zapytania, które z danego źródła zwraca maksymalnie trzy konkordancje na jednym ekranie wyników. Po wybraniu kryterium grupowania i określeniu maksymalnej liczby wyników, wyświetlone zostają co najwyżej trzy wystąpienia dopasowania w danej gazecie lub książce.

14.6. Metadane

W zaawansowanym formularzu zapytań możliwe jest także zawężenie wyszukiwania do wystąpień dopasowań w tekstach o zadanym typie funkcjonalnym, tytule lub też dacie publikacji. Domyślnie w polu metadanych musi wystąpić jedno lub więcej z podanych słów kluczowych, ale poprzedzając słowo kluczowe operatorem `AND`, możemy wymusić jego wystąpienie[9]. W polach metadanych można także stosować rozszerzenie ortograficzne oraz dowolnie zagnieżdżać

[9] Zob. `http://nkjp.uni.lodz.pl/?q=4c3axmp`.

Rysunek 14.7. Przykład grupowania konkordancji



warunki wystąpienia terminów. Na przykład wpisanie w polu *Tytuł źródła* warunku `gazeta AND (Lubuska OR Wrocławska)` ograniczy wyszukiwanie do tekstów z „Gazety Lubuskiej” oraz „Gazety Wrocławskiej”.

14.7. Wyrazy kontekstowe

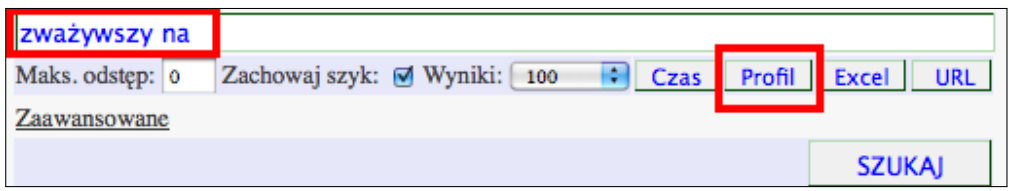
Pewne możliwości ujednoznaczniania wyników zapytania daje opcja określania wyrazów kontekstowych, które mogą lub nie powinny wystąpić w tym samym akapicie, którym znaleziono dopasowanie zapytania. Przypuśćmy, że szukamy wystąpień wyrazu *połączenie* w sensie *połączenie telefoniczne* i że chcemy automatycznie odsiać wszystkie konkordancje, które zawierają wyraz *kolejowy* albo frazę *z Internetem*. W tym celu wystarczy wpisać w polu *Wymagane wyrazy kontekstowe* zapytanie `zamiejscow* OR telefoniczn*`, a w polu *Niedopuszczalne wyrazy kontekstowe* zapytanie: `"z Internetem" OR kolejow*`. W zwróconych wynikach

powinny się w ten sposób znaleźć głównie wystąpienia rzeczownika *połączenie* w znaczeniu *połączenie telefoniczne*.

14.8. Analiza rejestru

Teksty NKJP opatrzone są informacją o typie funkcjonalnym, dzięki czemu możliwe jest sprawdzenie frekwencji występowania danego wyrazu lub frazy w różnych rejestrach języka. Aby wygenerować wykres słupkowy obrazujący frekwencję danego wyrazu lub frazy, wystarczy kliknąć przycisk *Profil* po wpisaniu zapytania. Na przykład po wpisaniu zapytania `zważywszy na` i kliknięciu przycisku *Profil* (rys. 14.8) wygenerowany zostaje wykres słupkowy podobny do wykresu ukazanego na rys. 14.9, z którego wynika, że fraza *zważywszy na...* pojawia się najczęściej w danych „quasi-mówionych", na przykład w sprawozdaniach stenograficznych Sejmu RP.

Rysunek 14.8. Generowanie profilu dla zapytania `zważywszy na`

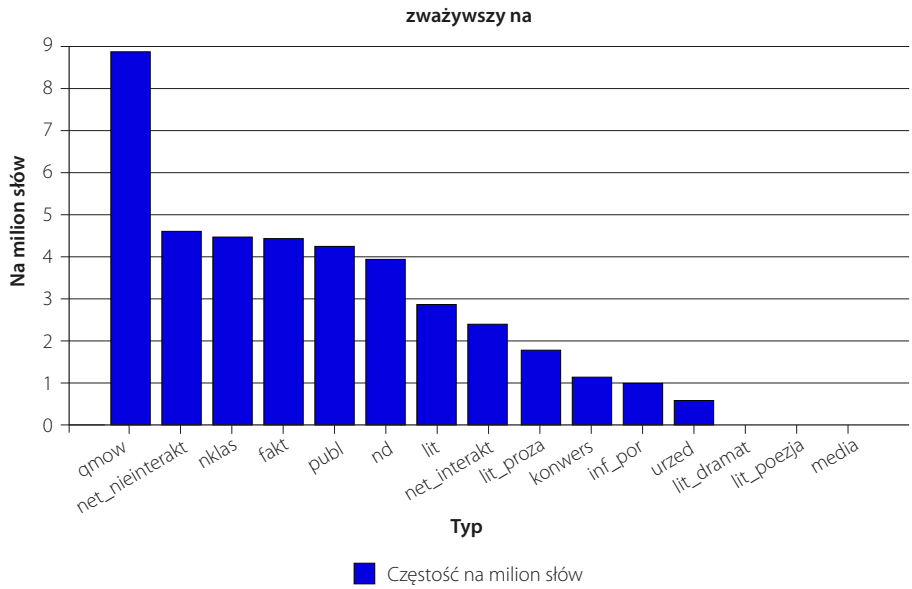


Poniżej wykresu częstości w różnych typach tekstów generowany jest wykres słupkowy częstości danego wyrazu lub frazy w „kanałach" publikacji uwzględnionych w taksonomii NKJP (rys. 14.10). Objaśnienia skrótów typów funkcjonalnych używanych w NKJP zawiera tab. 14.1.

14.9. Szeregi czasowe

Zasoby NKJP są bardzo zróżnicowane nie tylko ze względu na gatunek lub typ funkcjonalny tekstów, ale również z uwagi na czas ich powstania. Chociaż NKJP nie jest w zamierzeniu korpusem diachronicznym, w którym różne okresy czasu są równomiernie reprezentowane, to jednak dostępność informacji o dacie powstania lub pierwszej publikacji tekstu stwarza możliwości analizy frekwencji form językowych w zależności od czasu ich użycia. Analiza taka ukazuje, iż niektóre słowa, frazy, idiomy, nazwy własne i zwroty zyskują znacznie na popularności w krótkim czasie, odzwierciedlając tym samym nośność danego tematu w dyskursie publicznym.

Rysunek 14.9. Profil występowania dla zapytania `zważywszy na`



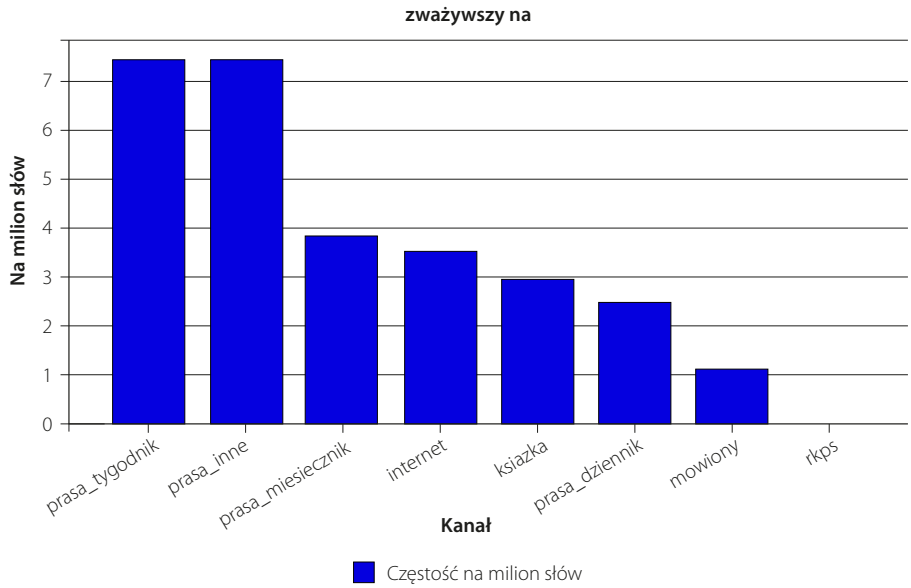
Wyszukiwarka PELCRA NKJP umożliwia wydobycie tego typu informacji o profilu diachronicznym słowa, lub frazy w bardzo prosty sposób. Po wpisaniu dowolnego zapytania w składni wyszukiwarki, należy kliknąć przycisk *Czas*. Po chwili poniżej formularza wyszukiwania powinien się pojawić wykres szeregu czasowego wraz z tabelą, na podstawie której został wygenerowany. Na przykład aby sprawdzić popularność słów *moherowy* lub *moher* we wszystkich odmianach w ostatnich 20 latach, należy wpisać zapytanie `moher**|moherowy**`, a następnie kliknąć przycisk *Czas* w formularzu zapytania (rys. 14.11).

Jak widać na wygenerowanym w ten sposób diagramie, popularność tych wyrazów wyraźnie wzrosła w latach 2005/2006 (rys. 14.12).

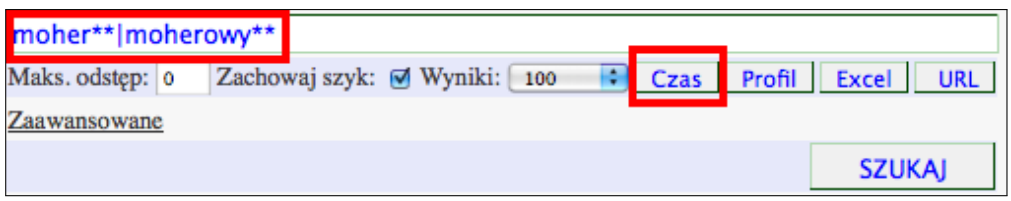
Odpowiednie zapytanie[10] o wystąpienia tych wyrazów po 2005 roku ukazuje przyczynę tego wzrostu frekwencji. Moher i moherowy beret nabrały w tym czasie metonimicznego znaczenia i zaczęły funkcjonować jako pejoratywne określenie pewnej grupy społecznej.

[10] Zob. `http://nkjp.uni.lodz.pl/?q=yzsxvo5`.

Rysunek 14.10. Wykres słupkowy częstości danego wyrazu lub frazy w „kanałach”



Rysunek 14.11. Zapytanie `moher**|moherowy**`



14.10. Pobieranie wyników w postaci arkuszy kalkulacyjnych

Widoczne na stronie wyniki wyszukiwania można pobrać z dodatkowymi metadanymi w postaci arkusza kalkulacyjnego, po kliknięciu przycisku *Excel*. Dzięki temu, użytkownik może dla własnych potrzeb sortować i edytować wyniki wyszukiwania. Arkusze z wynikami mają rozszerzenie .xml i należy je otwierać bezpośrednio z programu Microsoft Excel, Open Office lub Libre Office. Zeszyt wyników zawiera dwa arkusze. W arkuszu *Wyniki* można znaleźć konkordancje z podstawowymi metadanymi (rys. 14.13).

Tabela 14.1. Typy funkcjonalne tekstów w taksonomii NKJP

Skrót	Objaśnienie
publ	publicystyczne
net-interakt	internetowe interaktywne (np. fora, blogi z komentarzami, listy dyskusyjne)
net-nieinterakt	internetowe nieinteraktywne (np. strony domowe, blogi bez komentarzy)
nd	naukowo-dydaktyczne
qmow	quasi-mówione
fakt	literatura faktu
urzed	urzędowe
lit	literatura
inf-por	informacyjno-poradnikowe
nklas	inne
lit-poezja	poezja
media	mówione medialne
lit-proza	proza
konwers	mówione konwersacyjne
lit-dramat	dramat

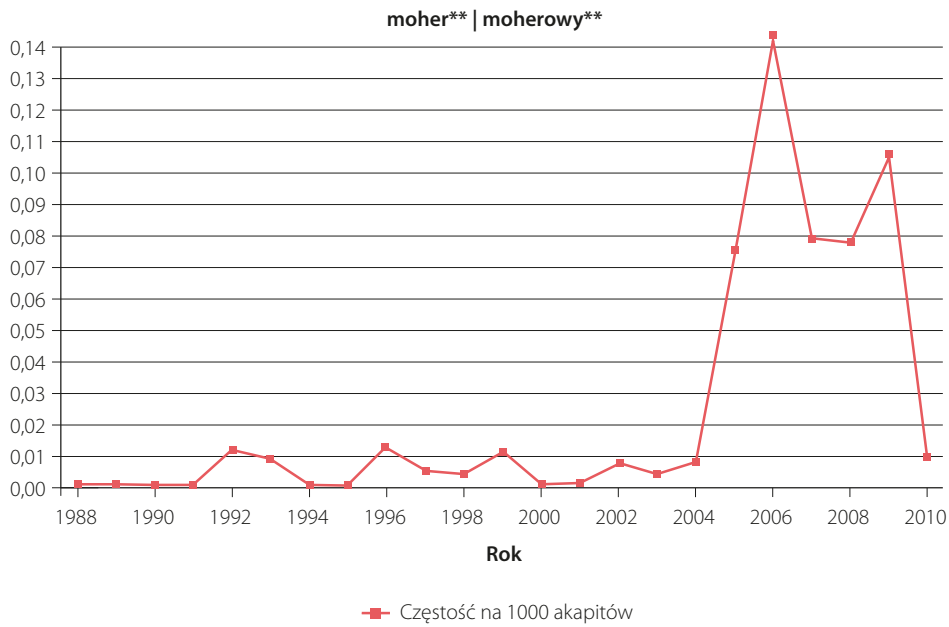
Warto zauważyć, że kolumna *left_word* zawiera słowo występujące bezpośrednio po lewej stronie dopasowania, dzięki czemu wyniki można sortować według lewego kontekstu. W arkuszu *Podsumowanie* znajdują się informacje o zapytaniu i zbiorze wyników.

14.11. Wyszukiwanie kolokacji

Korpusy językowe zawierają cenne informacje o łączliwości słów. Czasem typowe kolokacje danego wyrazu można wydobyć przez zwykłe posortowanie konkordancji po lewej lub prawej stronie. Badanie kolokacji przez sortowanie konkordancji może jednak okazać się kłopotliwe w wypadku często występujących słów. Na przykład różne odmiany rzeczownika *niebo* występują kilkanaście tysięcy razy w zrównoważonym podkorpusie NKJP. Ręczne przejrzenie wszystkich jego wystąpień w celu ustalenia najczęstszych kolokacji przymiotnikowych tworzonych z tym rzeczownikiem byłoby co najmniej niepraktyczne. Kolokator[11] to moduł automatycznej ekstrakcji kolokacji zaimplementowany w wyszukiwarce PELCRA NKJP, który znacznie ułatwia to zadanie. Narzędzie to jest dostępne w menu na górze głównej strony wyszukiwarki.

[11] Zob. `http://www.nkjp.uni.lodz.pl/collocations.jsp`.

Rysunek 14.12. Profil diachroniczny dla zapytania `moher**|moherowy**`



14.11.1. Ekstrakcja kolokacji pojedynczych wyrazów

Aby wyszukać lewostronne kolokacje przymiotnikowe rzeczownika *niebo* w różnych odmianach, należy najpierw sformułować odpowiednie zapytanie o ośrodek kolokacji, którym w tym wypadku jest wyraz *niebo*. W tym celu, używając opisanej powyżej składni, wpisujemy zapytanie `niebo**` do pola tekstowego ośrodka kolokacji, tak jak to ukazano na ilustracji poniżej. Dwie gwiazdki na końcu wyrazu oznaczają, że chodzi nam o wszystkie odmiany tego rzeczownika (rys. 14.14).

Kolejnym krokiem jest określenie kryteriów kolokacji. Ponieważ chcemy wyłuskać z korpusu kolokacje przymiotnikowe, z listy *Części mowy* wybieramy opcję *Przym./Imiesł.*, która uwzględnia przymiotniki i imiesłowy przymiotnikowe. Opcje *Kontekst z lewej* oraz *Kontekst z prawej* określają liczbę sąsiadujących z zadanym ośrodkiem kolokacji wyrazów, które mają być rozpatrywane jako część potencjalnych kolokacji.

Ze względu na złożoność obliczeniową ekstrakcji kolokacji, wyszukiwarka PELCRA chwilowo może jednorazowo w ciągu kilku sekund przeanalizować do 50 000 kontekstów wystąpień danego ośrodka kolokacji. Kolokacje wyrazów występujących w korpusie częściej niż 50 000 razy można wydobyć stopniowo, klikając przycisk *Następne*.

Rysunek 14.13. Wyniki w formacie arkusza kalkulacyjnego Excel (XML)

Sheets Charts SmartArt Graphics WordArt

<>	A	B	C	D	E	F	G
1	#	Left	Match	Right	pubDate	channel	title_mono
2	1	i starał się wyjść z	błędnego koła	prawd względnych i "	1996	#kanal_ksiazk	Legendy nowo
3	2	j jest analiza tego	błędnego koła	historycznego. Czerpa	1995	#kanal_ksiazk	Między Panem
4	3	ten sposób obwód	błędnego koła	zostaje zamknięty.	1997	#kanal_ksiazk	Finał klasyczne
5	4	wrócił do początku	błędnego koła	.	1957	#kanal_ksiazk	Kolumbowie —
6	5	az doświadczałem	błędnego koła	na skutek okazywane	2005	#kanal_ksiazk	Harar
7	6	nii jest przerwanie	błędnego koła	pesymizmu oraz ożyw	2002	#kanal_ksiazk	Choroba dyplo
8	7	zartykułować bez	błędnego koła	, ponieważ percepcja j	1974	#kanal_ksiazk	Obecność mitu
9	8	indukcji; krytyka	błędnego koła	zawartego w próbach	1974	#kanal_ksiazk	Obecność mitu
10	9	eguły indukcji bez	błędnego koła	. Ale argument taki ni	1974	#kanal_ksiazk	Obecność mitu
11	10	reguł, nie uniknie	błędnego koła	naturalizmu. Nadawar	1974	#kanal_ksiazk	Obecność mitu
12	11	zeczności (lub bez	błędnego koła	) rozumieć mit w jego	1974	#kanal_ksiazk	Obecność mitu
13	12	wodują powstanie	błędnego koła	, z którego wyjść moż	1996	#kanal_ksiazk	Podręcznik odr
14	13	tania mechanizmu	błędnego koła	. Używa się narkotyku	1974	#kanal_ksiazk	Melancholia

Rysunek 14.14. Zapytanie o ośrodek kolokacji

Po kliknięciu przycisku *Szukaj* należy odczekać kilkanaście sekund. Poniżej formularza zapytania powinna się ukazać tabela wyników, co ilustruje przedstawiony poniżej zrzut ekranu[12] (rys. 14.15).

Na górze tabeli wyników podana jest kolejno ogólna liczba wystąpień ośrodka kolokacji w korpusie, liczba przeanalizowanych kontekstów, oraz liczba potencjalnych kolokacji. Pierwsza kolumna tabeli wyników podaje liczbę porządkową kolokacji. W drugiej kolumnie wyświetlone są znormalizowane formy podstawowe kolokatów. Trzecia kolumna podaje konkretne kombinacje kolokacyjne dla wszystkich odmian formy podstawowej podanej w poprzedniej kolumnie, oraz

[12] Zob. `http://nkjp.uni.lodz.pl/?q=qg2gpv`.

Rysunek 14.15. Ekstrakcja kolokacji pojedynczych wyrazów – `niebo**`

#	Kolokacja	Pasujące współwystąpienia	Ogółem	Chi^2
1.	**rozgwieździć**	rozgwieżdżone___niebo *(30)*, rozgwieżdżonym___niebem *(13)*, rozg rozgwieżdżonemu___niebu *(1)*, niebem___rozgwieżdżonym *(1)*,	61	1,291,956.44
2.	**rozgwieżdżony**	rozgwieżdżone___niebo *(30)*, rozgwieżdżonym___niebem *(13)*, rozg rozgwieżdżonemu___niebu *(1)*, niebem___rozgwieżdżonym *(1)*,	61	1,291,956.44
3.	**bezchmurny**	bezchmurne___niebo *(23)*, bezchmurnym___niebie *(22)*, bezchmurne niebie___bezchmurnym *(3)*, nieba___bezchmurnego *(2)*, niebu___be	72	969,869.4
4.	**goły**	gołym___niebem *(202)*, gołe___niebo *(4)*, gołego___nieba *(2)*, niebi	209	539,774.91
5.	**wygwieździć**	wygwieżdżone___niebo *(7)*, wygwieżdżonym___niebem *(3)*, niebo_	14	494,209.52
6.	**gwiaździsty**	niebo___gwiaździste *(22)*, gwiaździste___niebo *(11)*, gwiaździstego_ nieba___gwiaździstego *(3)*, niebem___gwiaździstym *(3)*, gwiaździste	58	449,726.09
7.	**zachmurzyć**	zachmurzone___niebo *(17)*, zachmurzonym___niebie *(7)*, niebo___z zachmurzonego___nieba *(2)*, zachmurzonym___niebem *(1)*,	34	317,931.32
8.	**pochmurny**	pochmurne___niebo *(10)*, pochmurnego___nieba *(9)*, pochmurnym_ niebie___pochmurnym *(1)*, nieba___pochmurnego *(1)*,	29	201,454.9
9.	**jasne**	jasnego___nieba *(106)*, jasnym___niebie *(3)*, nieba___jasnym *(1)*,	110	189,997.74
10.	**siódmy**	siódmym___niebie *(86)*, siódme___niebo *(16)*, siódmego___nieba *(7*	109	105,407.01
11.	**błękitny**	błękitne___niebo *(27)*, błękitnego___nieba *(18)*, niebo___błękitne *(* błękitnemu___niebu *(1)*, nieba___błękitnymi *(1)*, niebu___błękitne_	65	95,748.27
12.	**granatowy**	granatowe___niebo *(13)*, granatowego___nieba *(10)*, granatowym___ *(2)*, niebie___granatowe *(1)*,	38	77,201.19

liczebności poszczególnych kombinacji. Po kliknięciu liczebności w osobnym oknie wyświetlane są konkordancje danej kombinacji. Pozwala to zweryfikować wyniki grupowania odmian do formy podstawowej. W czwartej kolumnie ukazano ogólną liczebność wszystkich form, która jest sumą form wszystkich kombinacji. Ostatnia kolumna podaje wartość chi kwadrat, która określa istotność statystyczną danej kolokacji. Właśnie według tej wartości sortowane są potencjalne kolokacje. Sortowanie kolokacji według zwykłej liczebności współwystąpień słów obniżałoby czytelność wyników, ze względu na dużą liczbę częstych słów, które tworzą z zadanym wyrazem związki składniowe, a nie kolokacyjne.

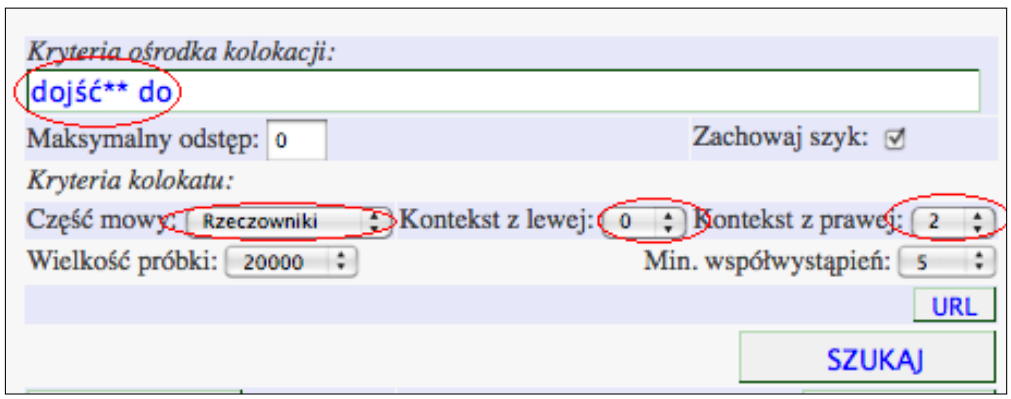
Jak widać, do typowych kolokacji rzeczownika *niebo* można zaliczyć takie frazy jak: *gołe niebo, rozgwieżdżone niebo, bezchmurne niebo, siódme niebo* itd., co chyba pozostaje w zgodzie z intuicją leksykalną użytkowników polszczyzny. Ciekawe są także informacje o preferencjach frazeologicznych wyłaniających się z liczebności niektórych form, np. *rozgwieżdżone*, *wygwieżdżone*, *gwieździste*, *gwiaździste niebo*.

Warto pamiętać, że wyszukiwarka nie zawsze jest w stanie rozszerzyć zapytanie o formy pokrewne morfologicznie. Na przykład jeżeli dla ośrodka kolokacji zdefiniowanego jako `VAT**` nie zwrócono żadnych wyników, to warto użyć zwykłego rozszerzenia ortograficznego, stosując zapytanie z jedną, a nie dwiema gwiazdkami, czyli `VAT*`.

14.11.2. Ekstrakcja złożonych kolokacji

Wyszukiwarka kolokacji umożliwia także badanie wielowyrazowych ośrodków kolokacji. Na przykład, aby wyszukać kolokacje występujące z czasownikiem *dojść* i przyimkiem *do*, można sformułować zapytanie ukazane na rys. 14.16[13].

Rysunek 14.16. Zapytanie dla kolokacji frazy `dojść** do`



Kontekst kolokacyjny ustawiono w tym przypadku na dwa słowa z prawej strony dopasowania. Jak widać na poniższym zrzucie ekranu, najbardziej istotne statystycznie kolokacje rzeczownikowe zwrócone przez powyższe zapytanie to między innymi *dojść do skutku/wniosku/porozumienia/przekonania* (rys. 14.17).

14.11.3. Jak rozumieć wartość chi kwadrat?

Potencjalne kolokacje są obecnie sortowane według wartości testu statystycznego chi kwadrat, który dość precyzyjnie określa jeden z aspektów łączliwości frazeologicznej.

Podana w ostatniej kolumnie tabeli wyników wartość chi kwadrat wyraża prawdopodobieństwo tego, że częstotliwość współwystępowania ośrodka kolokacji z danym wyrazem w korpusie nie jest przypadkowa. Dokładniej wyrażają to wartości prawdopodobieństwa przypisane do wartości chi kwadrat dla jednego stopnia swobody przedstawione w tab. 14.2.

Tabela 14.2. Typowe dla polszczyzny mówionej kombinacje segmentów wyrazowych

Chi kwadrat	2,706	3,841	5,024	6,635	10,828
Istotność statystyczna	0,90	0,95	0,975	0,99	0,999

[13] Zob. `http://nkjp.uni.lodz.pl/?q=oh3c5n`.

Rysunek 14.17. Ekstrakcja złożonych kolokacji – `dojść** do`

#	Kolokacja	Pasujące współwystąpienia	Ogółem	Chi^2
1.	**skutek**	dojdzie do___skutku *(144)*, doszła do___skutku *(105)*, doszedł do_ *(25)*, dojdą do___skutku *(14)*, doszłaby do___skutku *(11)*, doszedłb do___skutku *(1)*,	501	1,256,018.83
2.	**wniosek**	doszedł do___wniosku *(400)*, doszli do___wniosku *(265)*, doszedłe doszła do___wniosku *(125)*, doszłam do___wniosku *(72)*, dojdzie d do___wniosku *(25)*, dojdziemy do___wniosku *(23)*, dojdziesz do__ do___wniosku *(5)*, doszłaś do___wniosku *(5)*, doszliby do___wnio *(2)*, dojść do___wniosków *(1)*, doszedł do___wnioski *(1)*, doszła do do___wniosków *(1)*, doszli do___wniosków *(1)*, dojdź do___wnios	1711	1,003,276.97
3.	**porozumieć**	dojść do___porozumienia *(178)*, doszło do___porozumienia *(54)*, d do___porozumienia *(43)*, doszły do___porozumienia *(38)*, dojdzie do___porozumienia *(24)*, doszła do___porozumienia *(10)*, dojście d *(2)*, dojściu do___porozumienia *(2)*, doszedłem do___porozumienia do___porozumienia *(1)*, doszłaby do___porozumienia *(1)*,	529	487,055.59
4.	**zderzyć**	doszło do___zderzenia *(168)*, dojdzie do___zderzenia *(1)*, dojść do	170	273,322.09
5.	**kolizja**	doszło do___kolizji *(129)*, dojść do___kolizji *(2)*, dojdzie do___ko	133	146,936.19
6.	**szarpanina**	doszło do___szarpaniny *(27)*,	27	67,370.06
7.	**rękoczyn**	doszło do___rękoczynów *(12)*, doszłoby do___rękoczynów *(6)*, doj	24	58,876.79
8.	**tragedia**	doszło do___tragedii *(69)*, dojść do___tragedii *(32)*, dojdzie do___	121	53,072.95
9.	**władza**	dojściu do___władzy *(64)*, doszedł do___władzy *(60)*, dojście do__ *(20)*, doszli do___władzy *(19)*, dojdzie do___władzy *(19)*, dojść do do___władzy *(7)*, dojdziecie do___władzy *(6)*, doszło do___władzy doszliby do___władzy *(1)*, dojdę do___władzy *(1)*, dojście" do___	327	41,964.19
10.	**rozłam**	doszło do___rozłamu *(22)*, dojść do___rozłamu *(6)*, dojdzie do___	32	39,135.17

Jeżeli więc wartość testu chi kwadrat podana w tabeli wyników wynosi 10,828, to z matematycznego punktu widzenia istnieje tylko jedna szansa na tysiąc, że dane dwa wyrazy występują w zaobserwowanych kontekstach zupełnie przypadkowo. Innymi słowy, prawdopodobieństwo tego, że liczba współwystąpień wynika tylko i wyłącznie z ogólnej częstości występowania pojedynczych wyrazów wynosi 0,001.

W obecnej wersji wyszukiwarki wyświetlane są wyniki o liczebności współwystąpień >=5, oraz o wartości testu chi kwadrat >= 3,841. Oczywiście częstość współwystępowania wyrazów nie jest tylko funkcją ich łączliwości frazeologicznej i dlatego niektórych z współwystąpień wyrazów z wysoką wartością chi kwadrat nie można uznać za związki frazeologiczne.

14.12. Dostęp programistyczny

Wyszukiwarka PELCRA dla NKJP obsługuje także zapytania programistyczne przez protokół HTTP. Najlepiej to ilustruje skrypt napisany w języku Python[14]

[14] Pełna wersja skryptu dostępna jest pod adresem: `http://www.nkjp.uni.lodz.pl/getConcord.py.jsp`.

(wydr. 14.1). Skrypt ten wysyła zapytanie do serwera i otrzymuje wyniki konkordancji w prostym formacie XML (wydr. 14.2). Możliwe jest również automatyczne pobieranie wyników konkordancji we wspomnianym powyżej formacie Microsoft Excela (XML)[15].

Wydruk 14.1. Fragment skryptu napisany w języku Python

```
#coding=utf-8
import urllib
import random
servlet="http://nkjp.uni.lodz.pl/NKJPSpanSearchXML"
#Zapytanie w składni PELCRA NKJP
query="pleść** bzdura**"
#Maks. odstęp między słowami
span=2
#Zachowujemy szyk? true|false
preserve_order="false"
#Od którego wyniku zaczynamy?
offset=0
#od 1 do 5000 na raz. Wartości > 5000 są przycinane.
limit=50

#Inne parametry użyte w żądaniu HTTP poniżej... Zob.
http://www.nkjp.uni.lodz.pl/getConcord.py.jsp

params = urllib.urlencode({'query': query, 'offset': offset,
'span': span,'sort': sort, 'second_sort':'srodek', 'limit':
limit, 'groupBy':groupBy, 'groupByLimit':groupByLimit,
'preserve_order':preserve_order, 'dummystring':dummystring,
'sid':sid, 'm_date_from':m_date_from, 'm_date_to':m_date_to,
'm_styles':m_styles,
'm_channels':m_channels,'m_title_mono':m_title_mono,
'm_title_mono_NOT':m_title_mono_NOT,
'm_paragraphKWs_MUST':m_paragraphKWs_MUST,
'm_paragraphKWs_MUST_NOT':m_paragraphKWs_MUST_NOT})

f = urllib.urlopen(servlet, params)
print f.read()
```

Na razie nie stosujemy dodatkowych ograniczeń w automatycznym dostępie HTTP, ale mogą się one pojawić w wypadku nadużyć. Należy pamiętać, iż wszelkie formy komercyjnego wykorzystania wyszukiwarki wymagają uzyskania licencji od NKJP.

[15] Adres serwletu to `http://nkjp.uni.lodz.pl/NKJPSpanSearchExcelXML`.

Wydruk 14.2. Wyniki konkordancji w formacie XML

```
<?xml version="1.0" encoding="UTF-8"?>
  <results type="concordance">
  <index_size>1225343686</index_size>
  <!—Words in subcorpus—>
  <total_hits>808</total_hits>
  <query_time_s>0.034</query_time_s>
  <concordance>

  <line><count>1</count>
  <left><![CDATA[...że wreszcie udało mi się uciec choć na
  chwilę z tej okropnej Aleksandrii, brudnej i cuchnącej,
  gdzie nie można przejść dziesięciu jardów, nie napotkawszy
  tubylca załatwiającego się na skraju drogi. Cóż to za
  ]]></left><match><![CDATA[bzdury
  plecie]]></match><right><![CDATA[ się o Wschodzie! Czar
  Orientu! Slumsy w Wembley mają mniej więcej tyle samo
  uroku. A osławione piramidy wyglądają po prostu jak hałdy ż
  wiru... No cóż, to obowiązek żony oficera towarzyszyć mu
  wszędzie, ale przysięgam, że jeśli miano by go wysłać do
  Indii, zażądam, by wystąpił z armii. W końcu ja też
  zarabiam – powiedzmy sobie szczerze: moje dochody są
  większe niż jego apanaże...]]></right>
  <pubDate>1997</pubDate>
  <channel>#kanal_ksiazka</channel>
  <domain>#typ_lit</domain>
  <title_mono><![CDATA[Znak Anny ]]></title_mono>
  <title_a><![CDATA[]]></title_a>
  </line>

  </concordance>
  <!— itd. —>
</results>
```

14.13. Wyszukiwarka dla danych mówionych

W puli danych NKJP znajduje się obecnie ponad 2 miliony słów języka mówionego zarówno *medialnych*, jak też *konwersacyjnych*, które omówiono w oddzielnym rozdziale tego podręcznika. Wszystkie te dane można przeszukiwać w głównej wyszukiwarce NKJP, po wybraniu odpowiednich opcji typu i stylu tekstów.

Niemniej jednak, dzięki osobnej wyszukiwarce dla danych mówionych można wygodniej przeszukiwać, sortować i wyświetlać konwersacyjną część korpusu. Możliwe jest na przykład sortowanie konkordancji według wieku, wykształcenia lub też płci mówiącego[16].

14.14. Dalsze informacje

Aktualne informacje na temat wyszukiwarki podawane są na jej stronie pomocy, pod adresem: `http://nkjp.uni.lodz.pl/help.jsp`. Autor wyszukiwarki prosi o zgłaszanie uwag dotyczących działania wyszukiwarki na adres `piotr.pezik@gmail.com`.

[16] Zob. `http://www.nkjp.uni.lodz.pl/spoken.jsp`.

15

Słowa dnia

Anna Andrzejczuk i Marek Łaziński

Reprezentatywny zbiór tekstów Narodowego Korpusu Języka Polskiego pozwala wiarygodnie określić łączliwość leksykalną, a w konsekwencji typowe konotacje pojęciowe słów. Niezwykle istotną miarą ważności słowa i odpowiadającego mu pojęcia jest frekwencja w korpusie. Dopóki jednak nie zbudujemy listy frekwencyjnej na podstawie korpusu zrównoważonego tak szczegółowo jak *Słownik frekwencyjny polszczyzny współczesnej* (Kurcz i in. 1990), dane o porządku frekwencji słów współczesnej polszczyzny pozostaną wyrywkowe. W badaniach psycholingwistycznych wykorzystywano np. listę frekwencyjną zrównoważonego korpusu sieciowego PWN (`http://korpus.pwn.pl/`), ale jest to korpus jak na dzisiejsze warunki niewielki (20 mln słów). Jeśli chodzi o dane reprezentatywne dla polszczyzny ogólnej, to wciąż ostatnią metodycznie przygotowaną listą pozostaje wspomniany *Słownik frekwencyjny polszczyzny współczesnej*, wykorzystujący teksty z lat sześćdziesiątych.

Lista frekwencyjna, nawet oparta na zrównoważonym korpusie, jest statystycznie wiarygodna dla słów najczęstszych, które pojawiają się w rozmowach, prasie i książkach niezależnie od aktualnej mody, popularności utworów i produktów, sytuacji politycznej czy pór roku. Oprócz takiej listy, traktowanej jak dane referencyjne (tak był przez długie lata traktowany *Słownik frekwencyjny*), w codziennej prasie i rozmowach znajdujemy słowa, których aktualnie wysoka frekwencja jest wyraźnie zmienna w czasie. Informacje o słowach aktualnie najczęstszych: w danym dniu, tygodniu czy roku są istotne nie tylko z publicystycznego punktu widzenia. Bez takich danych nie sposób na przykład zdecydować, które nazwy własne czy nazwy produktów powinny trafić do kolejnych wydań słowników ortograficznych czy poradników językowych, a nawet do słowników ogólnych. Nazwy własne aktualnie najczęstsze powinny być też materiałem wyjściowym (ręcznie anotowanym), który służy do automatycznej ekstrakcji i lematyzacji innych nazw własnych.

Monitorowaniu częstości słów w prasie poświęcony jest podprojekt NKJP – Słowa dnia. Przez ten projekt chcemy uzupełnić standardowe funkcje korpusu referencyjnego o monitorowanie popularności poszczególnych słów.

Pomysł zaczerpnęliśmy z serwisu Wörter des Tages Uniwersytetu Lipskiego, który co dzień rano umieszcza na stronie internetowej `http://wortschatz.uni-leipzig.de/wort-des-tages/` słowa kluczowe dnia, wybrane na podstawie frekwencji w wydaniach internetowych kilkudziesięciu tytułów prasowych i sieciowych serwisów informacyjnych – por. Quasthoff i in. 2002. Lista Wörter des Tages – powstaje przez porównanie tekstów ściągniętych ze stron internetowych różnych gazet z tysiąckrotnie większym stałym (niezmienianym) korpusem referencyjnym prasy współczesnej. Słowa wybierane w formie podstawowej bez uwzględnienia odmiany, grupowane są automatycznie w kategorie semantyczne: 1. sportowcy, 2. sport (te dwie kategorie odpowiadają istotnej części zawartości prasy codziennej), 3. politycy, 4. organizacje, 5. wydarzenia, 6. hasła (niem. *Schlagwort*), 7. miejsca, 8. osoby ze świata kultury i nauki, 9. inne osoby. Dla każdego słowa można obejrzeć wybrane przykłady użycia, a także graf współwystępowania innych wyrazów w kontekście. Na stronach wydawnictwa Langenscheidt, z którym zespół Wörter des Tages współpracuje, publikowane są angielskie tłumaczenia słów z grupy hasła.

Polska lista słów prasy codziennej była publikowana w latach 2003–2006 na stronach „Rzeczpospolitej" oraz Korpusu Języka Polskiego PWN. Była przygotowywana w rytmie tygodniowym jako efekt współpracy redakcji „Rzeczpospolitej" online i zespołu Korpusu Języka Polskiego PWN (Łaziński i Szewczyk 2006). Lista gromadziła słowa w kategoriach: osoby, miejsca i mieszkańcy, inne nazwy własne (organizacje, firmy, wydarzenia) oraz rzeczowniki pospolite i inne słowa. Kategoryzacji dokonywano ręcznie. Korpus porównawczy dla słów danego tygodnia był tylko 12 razy większy, obejmował 3 miesiące. Był za to konsekwentnie aktualizowany, tzn. usuwano z niego teksty starsze niż 3 miesiące.

Podobne zasady zastosowano w zespole NKJP, zwiększono tylko liczbę tytułów prasowych oraz cykl czasowy. W latach 2009–2010 publikowane były zestawienia słów tygodnia prasy lokalnej spośród artykułów z kilkudziesięciu gazet zrzeszonych w Stowarzyszeniu Gazet Lokalnych.

Lista była zestawiana automatycznie – przez porównanie frekwencji względnej leksemów z uwzględnieniem dezambiguacji form wieloznacznych na podstawie listy zadanych proporcji obliczonych w ręcznie oznakowanym korpusie tekstowym – i co tydzień weryfikowana ręcznie przed publikacją.

Od marca 2010 funkcjonuje w ramach NKJP podstawowy serwis Słów dnia, oparty na frekwencji słów na stronach RSS czterech dzienników: „Dziennika. Gazety Prawnej", „Gazety Wyborczej", „Polski. The Times" oraz „Rzeczpospolitej".

Serwis działa nieprzerwanie i publikuje co rano słowa dnia poprzedniego. Autorem pierwszego programu do ekstrakcji słów, działającego do końca września 2010 roku był Daniel Janus, później zintegrowano działanie Słów dnia w NKJP z serwisem `www.frazeo.pl` autorstwa Piotra Pęzika. Wiosną 2011 roku w opracowywanie Słów dnia włączyli się językoznawcy z Instytutu Języka Polskiego UW, którzy będą to robić także po zakończeniu projektu NKJP.

Słowa dnia wybierane są na podstawie porównania frekwencji bezwzględnej i względnej w danym dniu oraz w całym poprzedzającym roku. Program proponuje listę słów w kolejności istotności statystycznej według testu G. Test ten podobnie jak test chi kwadrat stosowany w pierwszym programie Słów dnia, porządkuje słowa nie według kolejności ilorazów frekwencji w ostatnim dniu i w okresie porównawczym, lecz bierze dodatkowo pod uwagę bezwzględną liczbę wystąpień słowa.

W tabeli publikujemy z zasady pojedyncze słowa – ciągi znaków od spacji do spacji, a jeśli jakiś ciąg jest z leksykalnego punktu widzenia częścią większej całości: nazwy złożonej pospolitej lub własnej, dopisujemy brakujący fragment nazwy w rubryce *Komentarz*, np. (*katastrofa*) *smoleńska*. Opieramy się na teście istotności statystycznej, weryfikacja ręczna ogranicza się do usunięcia słów, których wysoka frekwencja nie ma związku z istotnością tematu w gazetach. Usuwamy np. nazwy dni tygodnia poprzedzających dzień, dla którego program wyszukuje słowa klucze. Jedyna ingerencja merytoryczna dotyczy wydarzeń sportowych. Program wyszukujący pomija w kanałach RSS dział sportowy (inaczej niż serwis Wörter des Tages, gdzie sport tworzy pierwszą kategorię), nazwiska czy słowa związane ze sportem trafią na listę wyłącznie wtedy, gdy znajdą się w gazetach na pierwszych stronach, a nie na stronach sportowych.

Słowa dnia prasy codziennej zaczęły się ukazywać na stronie NKJP w szczególnym czasie – w marcu 2010, a więc na krótko przed katastrofa smoleńską. Rzecz jasna, nazwiska, miejsca i pojęcia związane z katastrofą i pogrzebem ofiar dominowały w tabeli przez cały kwiecień. Słowa kluczowe przywołujące to wydarzenie pojawiają się regularnie do dziś. O niezwykłym miejscu tego wydarzenia w zbiorowej pamięci świadczy nie tylko frekwencja bezwzględna słów kluczowych, ale również ich różnorodność oraz fakt, że wiele różnych słów i grup słów odnosi się do tego samego wydarzenia czy kompleksu wydarzeń.

Wśród słów dnia związanych z katastrofą smoleńską część bezpośrednio dotyczyła tragedii (w pierwszych wiadomościach dotyczących tego wydarzenia): *Smoleńsk*, *katastrofa*, *tragedia*, *samolot*, *lot*, *lądować* oraz imiona i nazwiska ofiar i ich bliskich. Niedługo pojawiły się słowa związane z ostatnim pożegnaniem: *uroczystość*, *żałoba*, *pogrzeb*, *pochówek*, *trumna*, *Wawel*, ale także *rozkład jazdy* (pociągów do Krakowa).

W lipcu 2010 roku pojawiło się w tabeli słowo *krzyż*, które gościło tam nieprzerwanie przez dwa miesiące. Pamiętajmy, że aby słowo utrzymało się w tabeli długo, musiało co dzień mieć frekwencję większą niż w ciągu ostatnich trzech miesięcy, dlatego wysoka frekwencja słowa w okresie bezpośrednio je poprzedzającym utrudnia mu zdobycie pozycji kluczowej. Jeśli więc *krzyż* pojawiał się w gazetach przez dwa miesiące, to jego frekwencja w kolejnych dniach rosła, a przynajmniej nie malała. Kiedy już prasa przestała pisać o krzyżu spod Pałacu Prezydenckiego, słowo *krzyż* wróciło na naszą listę raz jeszcze pod koniec sierpnia 2010 roku – tym razem chodziło o usunięcie przez pracowników Tatrzańskiego Parku Narodowego krzyża ze szczytu Rysów, bo to wydarzenie wzbudziło u części publicystów podejrzenia o walkę z krzyżem w ogóle.

Kolejny etap rozwoju wypadków związanych z katastrofą smoleńską to publikacja raportu końcowego MAK. Wśród słów dnia oprócz: *MAK* (Mežgosudarstvennaja Aviacionnaja Komissija), *raport końcowy*, pojawiają się te, które dotyczą możliwych przyczyn katastrofy, których zabrakło w raporcie: *wieża*, *kontroler*, *rozmowa* oraz reakcji różnych osób na ten raport i różnych ocen tych reakcji: *skandal, Dolomity, Tusk, cynicznie, chłystek.*

Słowa kluczowe nie zawsze bezpośrednio nawiązują do konkretnej osoby, czy wydarzenia. Ilustrują to choćby tabele słów z okresu kampanii prezydenckiej. Zapełniała się ona co dzień nie nazwiskami kandydatów i nazwą ich partii (ich frekwencja nie rosła z dnia na dzień), ale słowami dawno niewidzianymi. W związku ze stosunkowo krótkim okresem porównawczym oraz porządkiem wybierania faworyzującym wysoką frekwencję bezwzględną nazwisko Edwarda Gierka, wspominanego jako patriota przez Jarosława Kaczyńskiego czy tragicznie zmarłej Barbary Blidy, z której mężem spotkał się Bronisław Komorowski, były w tabeli ważniejsze niż nazwy partii czy nazwiska liderów. Same nazwiska kandydatów trafiły na listę dopiero pod koniec kampanii wraz z ogłoszeniem wyników. Jako obiekt spotu w kampanii na listę trafiły również, tak mało się kojarzące z wyborami prezydenckimi słowa jak: *dopłaty* (do produkcji rolnej) czy *Afganistan* i problem misji wojskowej.

Wydarzenie tak tragiczne i wyjątkowe w dziejach państwa jak katastrofa smoleńska, nie jest najlepszą ilustracją typowych fluktuacji słów kluczowych w dziennikach w normalnych warunkach. Okazuje się, że wśród słów dotyczących wydarzeń, najczęściej się pojawiających i najdłużej goszczących w tabeli statystycznej, częściej znajdziemy wydarzenia związane z kataklizmami i wstrząsami politycznymi. W maju i czerwcu, a także jesienią 2010 roku w tabelach słów kluczowych królowały *powódź, zalać, woda, wał czy słup*, w marcu 2011 roku: *Japonia, tsunami, trzęsienie* (*ziemi*), *elektrownia*, *atomowa*, *Fukushima, reaktor, promieniowanie, radiacja.* Zimą i wiosną roku 2011 (kiedy powstawał ten tekst) najwięcej słów klu-

czowych związanych było z zamieszkami w Afryce: *Egipt*, *egipski*, *Egipcjanin*, *Kair*, *prezydent*, (*Hosni*) *Muburak*, *zamieszki*, *antyrządowy*, *Tahir*, *zwolennik*, *manifestant*, *demonstrant*, *ustąpić*, *Szafik*, (*El*) *Baradei*, *Libia, libijski, Unuf, Zintan, Bengazi, Zawija, Trypolis, Kaddafi* (*Kadafi*), *przeciwlotniczy* (ogień), *myśliwiec, samolot, Tomahawk, powstaniec, partyzant, reżim, zabity*.

Nazwiska pojawiające się najczęściej w tabelach w roku 2010 należą do polityków: *Bronisław Komorowski* – 43 razy, *Jarosław Kaczyński* – 20, *Janusz Palikot* – 9 razy, *Grzegorz Napieralski* – 8 razy. (Nie uwzględniamy słów, które pośrednio nawiązują do danej postaci, np. *immunitet* Jarosława Kaczyńskiego, *mandat* Janusza Palikota czy *polowanie* Bronisława Komorowskiego). Na drugim miejscu, choć bardzo daleko za czołówką, znajdują się nazwiska sportowców: *Adam Małysz – medal, benefis, skok, Zakopane* – oraz *Robert Kubica – wypadek, dłoń, śpiączka, wybudzić, operacja, łokieć* (na liście świadomie ograniczamy udział informacji sportowych). Dopiero na końcu pojawiają się nazwiska związane z kulturą – pisarzy, aktorów, naukowców, piosenkarzy. Także te nazwiska stają się jednak głośne w prasie codziennej, często w związku z wydarzeniami pozaartystycznymi, takimi jak proces Romana Polańskiego, a często – dopiero po śmierci – *Krzysztof Kolberger*, *Elizabeth Taylor*, *Karin Stanek*.

Spośród słów oznaczających wydarzenia część odnosi się do zjawisk grających ważną rolę w polskim życiu publicznym przez dłuższy okres: O*FE*, *VAT*, *autostrady*, wysoka frekwencja innych słów trwa krótko i jest związana z pojedynczym wydarzeniem lub przejściowymi problemami, np. *chaos* (w *PKP*). Są również słowa kluczowe cykliczne, związane ze świętami i wydarzeniami powtarzającymi się co rok: *Święta, Boże Narodzenie*, *choinka, Wigilia, Sylwester*, *WOŚP* (Wielka Orkiestra Świątecznej Pomocy), *wolontariusz, finał* lub z rocznicami wydarzeń z przeszłości: *stan wojenny*, *Jaruzelski, ZOMO*, (kopalnia) *Wujek*.

Dalsze badanie słów w takich grupach pozwoli sprawdzić, jak często naszą uwagę zajmują informacje czy obiekty zupełnie nowe, jak często przywołujemy obiekty z pamięci biernej. Listę słów kluczowych można w tym celu potraktować jak tekst i szukać w nim słów najczęstszych, czyli słów superkluczowych charakterystycznych dla różnych dłuższych okresów lub dla kręgów tematycznych (*key key words* – por. Scott 2007). Takie badania są na razie w planach naszego zespołu.

Dla użytkowników Narodowego Korpusu Języka Polskiego Słowa dnia mają być prostą prezentacją pojęć i obiektów, które zapisywane są w zbiorowej pamięci najczęściej, co nie znaczy – na najdłuższy czas. Widoczne są słowa, których używamy często po raz pierwszy lub przypominamy sobie o nich sobie po dłuższym braku zainteresowania.

Część V

Zastosowania

16

NKJP w oczach leksykografa

Mirosław Bańko

W niniejszym artykule chcielibyśmy spojrzeć na NKJP oczami leksykografa. Omówimy tradycje wykorzystania danych źródłowych w pracy nad słownikami, a także znaczenie korpusów językowych dla współczesnych słowników. Przypomnimy niektóre polskie słowniki oparte na obszernych zbiorach danych źródłowych, tj. kartotekach i korpusach. Dotkniemy też społecznych warunków funkcjonowania leksykografii, w tym okoliczności, które ograniczają wykorzystanie korpusu, nawet gdy jest on łatwo dostępny.

Artykuł jest z założenia mało techniczny: nie mówi, jak korzystać z NKJP w pracy nad słownikami, nie pokazuje przykładowych kwerend w korpusie. Sposób pracy z korpusem zależy bowiem od rodzaju słownika, a jedyny obecnie słownik, o którym wiadomo, że powstaje na podstawie NKJP – mianowicie *Wielki słownik języka polskiego* (Żmigrodzki i in. 2007) – nie upoważnia do uogólnień. Jedno można wszak powiedzieć na samym wstępie: dzięki zaawansowanym narzędziom, w jakie został wyposażony NKJP, leksykografowie zyskali potężne wsparcie w ich pracy.

16.1. Kartoteki w pracy nad słownikami

Niewielki słownik można opracować na podstawie większego słownika, całkiem duży słownik można zestawić przez kompilację wcześniej wydanych leksykonów, ale w ten sposób – jeśli nawet pominąć kwestie prawne i etyczne – nie przygotuje się istotnie nowego dzieła i nie opisze nowych wyrazów. Tych ostatnich trzeba więc szukać w tekstach, co ma ten plus, że teksty informują o użyciu wyrazu, pośrednio o jego znaczeniu, a więc nie tylko dowodzą, że wyraz istnieje, ale też ułatwiają jego opis. Poza tym teksty można cytować, a cytaty, zwłaszcza z dzieł znanych autorów, podnoszą prestiż słownika, czynią go bardziej wiarygodnym i w końcu – choć to rzecz subiektywna – przyjemniejszym w odbiorze. Nic dziwnego więc,

że autorzy słowników od dawna podstawą swojej pracy czynili teksty źródłowe: z nich czerpali wyrazy, przykłady ich użycia, a nawet objaśnienia.

Na przykład *Thesaurus Polono-Latino-Graecus* Knapiusza (1621) zawiera odesłania do kilkudziesięciu autorów polskich, których autorytetem twórca słownika chciał uzasadnić obecność w nim niektórych rzadkich wyrazów. *Nowy dykcjonarz, to jest mownik polsko-niemiecko-francuski* (1764) Troca został oparty na rozleglejszym zasobie źródeł, lecz trudno o tym wiele powiedzieć, gdyż autor nie wymienił cytowanych dzieł, zastąpiwszy ich wykaz odesłaniem do katalogu Biblioteki Załuskich. Kilkadziesiąt utworów cytuje, i to nawet z dokładnością do strony, Bandtkie w *Słowniku dokładnym języka polskiego i niemieckiego* (1806). Przełomem w leksykografii polskiej stał się jednak dopiero *Słownik języka polskiego* (1807–1814) Lindego, oparty na ponad 850 tekstach źródłowych, reprezentujących różne odmiany piśmiennictwa polskiego od XVI po koniec XVIII wieku, nie wyłączając literatury popularnej ani tekstów fachowych.

Warto pamiętać, że Linde nie tylko napisał pierwszy polski słownik jednojęzyczny, ale też zamierzał opracować bibliografię polską (Ptaszyk 2007: 19 i nn.). Z tego drugiego zadania wprawdzie zrezygnował, ale jako bibliotekarz hr. Józefa M. Ossolińskiego w Wiedniu opisał bogaty księgozbiór swojego mecenasa (pozostawiony przez Lindego rękopis liczy 22 tomy), więcej nawet – księgozbiór ten aktywnie pomnażał, podejmując podróże po Galicji, z których przywoził książki. W liście do Ossolińskiego donosił raz, że z klasztoru bernardynów, gdzie „ni perswazye, ni pieniądze, ani zamiany, ani rewersa, ani pogróżki nic nie pomogą", książki musiał „smykać", tj. kraść (Matuszczyk 2006: 26). Takie zaangażowanie wynikało z przekonania Lindego, że o wartości słownika decyduje podstawa źródłowa, a kompetencja autora schodzi na dalszy plan. Linde-leksykograf i Linde-bibliotekarz dobrze uzupełniali się w pracy, można powiedzieć wręcz, że z pobudek leksykograficznych stał się Linde twórcą pierwszego polskiego „korpusu" językowego, który wykorzystał w pracy nad słownikiem.

O technice pracy Lindego z owym „korpusem" świadczą zachowane w zbiorach Ossolineum egzemplarze zakreślone przezeń czerwonym ołówkiem. Zaznaczone w ten sposób fragmenty Linde lub zatrudniony przezeń kopista przenosił na kartki, te zaś – według relacji generała Józefa Załuskiego, który odwiedził Lindego w Wiedniu – nasz leksykograf trzymał w koszykach, które wypełniały cały pokój (Załuski 1976: 44). Późniejsza relacja Fryderyka Skarbka, ucznia Liceum Warszawskiego w czasie, gdy Linde był jego dyrektorem, przynosi nieco inny obraz: „miał [Linde] jeden pokój cały zastawiony szufladami z przegrodami do zecerskich kaszt podobnymi, w których znajdowały się porządkiem liter kartki z wypisami autorów lub z definicjami wyrazów do Słownika wchodzących" (za Michalskim 1961: 5). Jak widać, twórca

polskiego słownika narodowego doskonalił swój warsztat i od pewnego czasu posługiwał się dobrze nam znaną kartoteką – podobnie jak całe pokolenia leksykografów po nim, aż do niedawna, a w pewnej mierze do dziś. Ze wstępu do jego słownika (s. IV) wynika, że do kartoteki tej wkładał nawet wyrazy niepoświadczone w druku, zbierane „po warsztatach", a więc żargonowe. Dziś należałoby ich szukać podobnie, tyle że z dyktafonem w ręce, bo nie można założyć, że wszystkie współczesne „warsztaty" mają już swoje witryny internetowe, pozwalające poznać ich słownictwo bez wychodzenia z domu.

Dwa następne znaczące słowniki polskie – tzw. wileński (1861) i warszawski (1900–1927) – też miały swoje „korpusy", lecz ich zawartości możemy się tylko domyślać, gdyż w żadnym ze słowników nie podano wykazu ekscerpowanych dzieł, a przykładów nie zlokalizowano w tekstach źródłowych (w słowniku warszawskim są tylko skróty nazwisk cytowanych autorów). Można powiedzieć za to, że sam słownik warszawski miał w myśl swojego pomysłodawcy i współtwórcy, Jana Karłowicza, pełnić taką funkcję, jaką dziś mają korpusy językowe, tzn. „być bezstronnym, nieuprzedzonym żadną teoryją gramatyczną lub purystyczną, ścisłym inwentarzem języka, dokładną i obszérną skarbnicą jego zasobów, skarbnicą, że tak powiem, bezwyznaniową, jedném słowem materyjałem, z którego dopiéro językoznawcy spostrzeżenia swe i wnioski snuć będą" (Karłowicz 1876: XVIII–XIX). Ten programowy anormatywizm został w słowniku warszawskim złagodzony, gdyż wyrazy będące w użyciu, a oceniane jako niepoprawne wprawdzie w nim przytoczono, ale poprzedzono ostrzegawczym wykrzyknikiem. Można uznać Karłowicza za wczesnego rzecznika postawy, którą dziś określamy jako normatywizm opisowy (Bańko 2001: 45–49), a której wdrożenie zapowiedzieli inicjatorzy *Wielkiego słownika języka polskiego* PAN (Żmigrodzki i in. 2007: 10).

Największą kartotekę słownikową w dziejach leksykografii polskiej zgromadzono w trakcie prac nad *Słownikiem polszczyzny XVI wieku* – liczy ona ponad 8 milionów kart (`http://www.ibl.waw.pl/index.php?strona=206`). Prawdopodobnie rekord ten nie zostanie nigdy pobity, gdyż nowsze słowniki będą oparte na zasobach komputerowych. Dużą kartotekę miał także *Słownik języka polskiego* pod red. Witolda Doroszewskiego, jej objętość były kierownik redakcji, która przygotowała słownik, szacuje na 6 milionów kart (Szkiłądź 1997: 327). Późniejszy i mniejszy, oparty na poprzedniku *Słownik języka polskiego* pod red. Mieczysława Szymczaka musiał się zadowolić kartoteką liczącą „tylko" około miliona cytatów. Dziś pierwsza z nich znajduje się na Wydziale Polonistyki Uniwersytetu Warszawskiego, druga na Wydziale Filologicznym Uniwersytetu Opolskiego. Obie mogą służyć badaczom i studentom, po części jako źródła informacji, po części zaś jako zabytki kultury piśmiennej minionej epoki.

16.2. Znaczenie korpusu dla leksykografii

Słowniki służą różnym celom, co wynika z ich ogromnej różnorodności (zob. Bańko 2010). Nie można więc utrzymywać, że korpus językowy jest jednakowo potrzebny w pracy nad każdym słownikiem. Nie jest też tak, aby jeden korpus, np. NKJP, mógł być podstawą wszystkich słowników danego języka. Dane źródłowe w leksykografii, tak jak w ogóle w nauce, trzeba gromadzić pod kątem konkretnego projektu, ale korpus narodowy może to zadanie ułatwić, może też umożliwić wstępną weryfikację hipotez.

Jeszcze niedawno, bo w latach 90. XX wieku, niektórzy polscy językoznawcy byli skłonni pomniejszać rolę korpusu w badaniach nad językiem, a także w pracy leksykografa. Twierdzono wówczas, że żaden korpus nie odzwierciedli kompetencji językowej rodzimego użytkownika języka, że w korpusie widać przede wszystkim zjawiska seryjne, o których wiadomo i bez korpusu, że w popularnym słowniku przykłady wymyślane przez redaktorów lepiej pełnią funkcję ilustracyjną niż zdania cytowane ze źródeł, zwykle zbyt skomplikowane i za długie (zob. Bańko 2001: 26–29). Nieufność do korpusu jako źródła danych cechowała zresztą niemal całą lingwistykę zachodnią, na której silne piętno odcisnął generatywizm, z założenia introspekcyjny, niechętny badaniom empirycznym. Stubbs (1996: 22–50), który analizował prace autorów anglosaskich, dowodzi jednak, że antyempiryzm Noama Chomsky'ego od początku kontrastował z empirycystycznym podejściem takich badaczy, jak John Firth, Michael Halliday, John Sinclair czy William Labov. Głosili oni, że język jest przede wszystkim zjawiskiem społecznym, a nie psychologicznym czy biologicznym, że badanie go polega na analizie autentycznych tekstów, a nie izolowanych, układanych przez badacza zdań, że zjawiskiem nie mniej godnym uwagi niż wyolbrzymiana przez Chomsky'ego kreatywność języka jest rutynowość w jego użyciu, „śmiertelnie nudna powtarzalność" (określenie Dwighta Bolingera 1965: 570, *deadly repetitiousness of language*). Dziś, z perspektywy czasu widać, że stanowisko przeciwników badań korpusowych było w dużej mierze próbą racjonalizacji ich postaw metodologicznych w czasach, gdy dostatecznie duże korpusy nie były łatwo dostępne. Zjawisko takie znane jest w psychologii pod nazwą „słodkie cytryny" i objaśniane następująco: ktoś, kto doświadcza nieprzyjemnej sytuacji, a nic nie może na to poradzić, jest skłonny dla własnego komfortu psychicznego twierdzić, że jego sytuacja nie jest zła, może nawet jest wyśmienita, co więcej – jest gotów w to uwierzyć.

Ciekawe, że choć grupa wpływowych lingwistów zachodnich zdołała przekonać swoich kolegów, że cytryny są słodkie, tzn. niedostępność dużych korpusów nie jest niczym złym, wśród przekonanych niewielu było leksykografów. Leksykografia bowiem – może ze względu na swój tradycjonalizm (czasem oceniany

krytycznie jako zapoźnienie metodologiczne), a może po prostu ze względu na charakter słowników – hołdowała wciąż empiryzmowi. Słownikarze w połowie XX wieku, jak Linde 150 lat wcześniej, wciąż ekscerpowali teksty, tzn. zakreślali w nich fragmenty, przepisywali je na karteczki, a karteczki układali pracowicie w kartotekach. Praca nad hasłem słownikowym przypominała wtedy stawianie pasjansa: kartki trzeba było rozkładać na biurku, dzieląc je na grupy według dających się zaobserwować typów użyć, docelowo zazwyczaj znaczeń, przy czym pierwsza próba często kończyła się wyodrębnieniem zbyt wielu grup, trzeba było więc następnie je scalać. W wypadku częstych wyrazów należało przejrzeć setki kartek, co zabierało tyle czasu, że ów słownikowy pasjans ciągnął się nawet dłużej niż niejeden karciany. Co gorsza, ponieważ praca nad hasłem wymagała często wielokrotnego grupowania cytatów według różnych kryteriów, pasjans trzeba było powtarzać.

Nic dziwnego, że redaktorzy słowników stosunkowo wcześnie – wcześniej od wielu lingwistów – dostrzegli pożytki z komputeryzacji prac słownikowych, a w szczególności z gromadzenia materiałów źródłowych w formie elektronicznej. Początkowo myślano tylko o komputeryzacji tradycyjnych kartotek, np. Szkiłądź (1997: 336) wspomina, że Redakcja Słowników Języka Polskiego PWN, wówczas pod jego kierownictwem, zabiegała o to już w 1980 r. W tym samym roku wydawnictwo Collins we współpracy z uniwersytetem w Birmingham przystąpiło do pracy nad wielomilionowym korpusem angielszczyzny, który stał się podstawą pierwszego w świecie słownika opartego na korpusie komputerowym, wydanego w 1987 r. pod tytułem *Collins Cobuild English Langauge Dictionary* (Sinclair 1987). W ten sposób dokumentacja źródłowa, poprzednio zajmująca setki lub tysiące szuflad w kartotekach, stała się dostępna bez wstawania od biurka, w ilości niewyobrażalnej, w dodatku podatna na automatyczną selekcję i porządkowanie w formie wygodnej dla leksykografa.

Choć korpusy do celów badawczych powstawały po obu stronach Atlantyku od lat 60. XX wieku, były za małe na potrzeby leksykografii. Dopiero wyżej wymieniony słownik Collinsa pokazał, że praktyczne wykorzystanie korpusu w leksykografii komercyjnej jest możliwe, co więcej – że korpus jest szczególnie użyteczny w pracy nad tzw. *learner's dictionaries*, czyli słownikami dla cudzoziemców (dosłownie: uczniowskimi, ale wśród takich „uczniów" są też osoby dorosłe). Wydaje się, że nie tylko rosnąca rola języka angielskiego w świecie, ale też dostępność korpusów językowych sprawiła, że obecnie na rynku *learner's dictionaries* konkuruje pięciu dużych wydawców, którzy oferują odbiorcom innowacyjne słowniki, dostarczające im wiarygodnych i szczegółowych informacji. To dzięki korpusom i słownikom korpusowym dziś dobrze rozumiemy, że są takie aspekty użycia wyrazu, których bez wglądu w duży korpus po prostu nie

da się dobrze opisać, gdyż w ich wypadku intuicja zawodzi. Idzie tu zwłaszcza o względną frekwencję wariantów, o wybór typowych kolokacji, o mniej seryjne własności gramatyczne opisywanych jednostek, a także o informację normatywną – o ile oczywiście leksykograf zdecyduje się odrzucić normatywizm wyłącznie selekcyjny, utrwalający tradycyjne poglądy i przesądy, a w zamian wdrożyć tzw. normatywizm opisowy, oparty w równej mierze na tradycji, co na danych z korpusu (zob. Bańko 2001: 37–49, gdzie mowa o tych i innych zastosowaniach korpusu w pracy nad słownikami).

Pierwszym w leksykografii polskiej słownikiem w znacznej mierze korpusowym – jeśli nie liczyć list frekwencyjnych wydanych w latach 1974–1977 pt. *Słownictwo współczesnego języka polskiego* i scalonych dopiero w r. 1990 w publikacji *Słownik frekwencyjny polszczyzny współczesnej* – był chyba *Inny słownik języka polskiego* PWN. Prace nad nim rozpoczęto, gdy Korpus Języka Polskiego PWN jeszcze nie istniał, toteż rzeczowniki i część przymiotników opracowano na podstawie kartoteki sporządzonej jeszcze na potrzeby *Słownika języka polskiego* pod red. M. Szymczaka i uzupełnianej do początku lat 90. Natomiast pozostałe przymiotniki, a prócz nich czasowniki, spójniki, przyimki, partykuły i zaimki opisano w całości na podstawie korpusu PWN, wówczas obejmującego ok. 25 milionów słów w różnych publikacjach książkowych i prasowych, a także w tekstach mówionych. Zróżnicowanie korpusu i zachowanie względnej równowagi między różnymi rodzajami tekstów było dla PWN od samego początku priorytetem.

W następnych latach w PWN korpus służył jako pomocnicze narzędzie w pracy nad wieloma słownikami, zarówno polskojęzycznymi, jak i przekładowymi, m.in. nad *Wielkim słownikiem angielsko-polskim* (2002) i *Wielkim słownkiem polsko-angielskim* (2004). Spośród nich jednak tylko jeden można bez wahania nazwać korpusowym – *Słownik dobrego stylu* z roku 2006. W istocie jest to słownik kolokacji, czyli zgodnie z polską tradycją nazewniczą – związków frazeologicznych łączliwych, leżących na pograniczu stałych i luźnych połączeń wyrazowych. Wydawnictwo tak dalece obawiało się słowa *kolokacje* w tytule, że publikację wydano pod tytułem nawiązującym do podobnych słowników niemieckich, być może bardziej obiecującym, ale mniej adekwatnym, gdyż zjawiska w niej opisane nie mają związku z kategorią stylu. Zarówno wybór haseł, jak i szczegółowe decyzje redakcyjne podporządkowano tu analizie statystycznej ok. 50 milionów słów z korpusu PWN. Użytkownicy NKJP mogą obecnie porównać kolokacje *Słownika dobrego stylu* z kolokacjami dostarczanymi przez wyszukiwarkę PELCRA (i oczywiście mogą generować kolokacje dla wyrazów, które w słowniku nie figurują jako hasłowe).

Dostępny od niedawna NKJP służy *Wielkiemu słownikowi języka polskiego*, nad którym czuwa Instytut Języka Polskiego PAN (zob. `http://wsjp.pl/`), oba

projekty wystartowały zresztą w tym samym czasie. Ponieważ jednak przykłady z NKJP są przytaczane w wielu innych pracach, trudno przypuszczać, aby nie było wśród nich słowników, które niedługo zostaną wydane. O jednym z nich już wiadomo: Wydawnictwa Uniwersytetu Warszawskiego mają w planie słownik fraz odimiennych pt. *Ludzie i miejsca w języku* (autorami są Maciej Czeszewski i Katarzyna Foremniak), który zamierzają anonsować jako pierwszy opracowany na podstawie NKJP.

16.3. Sam korpus nie wystarczy

Poza PWN żaden inny polski wydawca słowników nie chwali się na okładce ani w innym miejscu posiadaniem korpusu językowego. Jednocześnie oferta słowników na rynku jest duża, samych tylko słowników ortograficznych sprzedawcy oferują ponad sto, a słowników wyrazów obcych – kilkadziesiąt. Jeśli wydawcy i autorzy popularnych słowników nie przyznają się do korzystania z korpusów, to znaczy, że publikacje swoje opierają głównie na wcześniejszych słownikach, co oczywiście ogranicza innowacyjność nowych tytułów i może rodzić problemy z prawem autorskim. Kto by jednak przypuszczał, że obecnie, gdy dostępny jest NKJP, sytuacja zmieni się diametralnie i na naszych oczach zaczną powstawać doskonałe, z roku na rok coraz lepsze słowniki, ten jednak byłby w błędzie. Samo istnienie korpusu nie wystarczy, aby autorzy i wydawcy słowników zaczęli z niego korzystać. Muszą oni w tym dostrzec dla siebie jakąś korzyść – prywatni wydawcy przede wszystkim finansową – a nabywcy słowników muszą mieć wystarczającą świadomość leksykograficzną, aby ich do tego zachęcić bądź zmusić. W szczególności chcielibyśmy tu zwrócić uwagę na kilka okoliczności, które hamują rozwój leksykografii korpusowej w Polsce.

Po pierwsze, redagowanie haseł na podstawie korpusu jest wprawdzie szybsze i tańsze niż opracowywanie ich na podstawie tradycyjnej kartoteki, lecz jeszcze szybsze i tańsze jest kopiowanie informacji z innych słowników. Dopóty, dopóki jakiś niekorpusowy słownik sprzedaje się dobrze, jego wydawca nie ma powodu finansować prac, których celem byłoby zastąpienie go słownikiem opartym na korpusie.

Po drugie, w leksykografii, niezależnie od uwarunkowań finansowych, tradycja zawsze liczyła się bardziej niż innowacyjność, a dokonania przodków – nawet jeśli w dyskusjach akademickich podważane – w praktyce były rozwijane i kontynuowane. Słowniki są zachowawcze i kumulatywne, ponieważ tego od nich oczekuje społeczeństwo.

Po trzecie, większość użytkowników słowników nie ma wystarczającej świadomości leksykograficznej, aby odróżnić wybitny słownik od przeciętnego,

a także by wybrać słownik odpowiedni do swoich potrzeb. Wprawdzie nie przeprowadzono w Polsce na dużą skalę badań nad używaniem słowników (z wyjątkiem słowników dwujęzycznych, których użyciu poświęcono wiele prac), ale ankiety wykonane w środowisku uczniów, nauczycieli i studentów – a więc osób, które można by podejrzewać o większą niż przeciętna świadomość leksykograficzną – pokazały, jak bardzo wiedza i samowiedza respondentów jest ograniczona (Piper 2003, Żmigrodzki i in. 2005).

Po czwarte, nabywcy i użytkownicy słowników na ogół niewiele wiedzą o korpusach językowych, a postulat, by korpus uczynić podstawą ocen normatywnych w opisie języka spotyka się z nieufnością także wśród językoznawców. Twórcy korpusów tym większą wagę powinni przykładać do ich struktury. Źle zrównoważony korpus może bowiem poderwać zaufanie do korpusu jako narzędzia w ogóle.

Wydaje się, że aby leksykografia szerzej otworzyła się na dobrodziejstwo korpusów językowych, trzeba z jednej strony doskonalić i upowszechniać przyjazne w obsłudze narzędzia usprawniające przetwarzanie dużych ilości danych zawartych w korpusie (takie jak Sketch Engine, zob. Kilgarriff i in. 2008, zob. też `http://www.sketchengine.co.uk/`). Z drugiej strony zaś należy pracować nad świadomością leksykograficzną użytkowników słowników, aby ci – przede wszystkim przez swoje decyzje nabywcze – wywierali nacisk na wydawców i w ten sposób nakłaniali ich do korzystania z nowoczesnej technologii. Edukację leksykograficzną powinno prowadzić się już w szkole, zapoznając uczniów ze słownikiem jako, po pierwsze, praktycznym narzędziem, po drugie zaś – zwierciadłem epoki i kultury.

17

Zastosowanie korpusów w badaniu gramatyki

Rafał L. Górski

O ile zastosowania korpusów w leksykologii, leksykografii czy badaniu kolokacji wydają się oczywiste, o tyle mniej oczywista jest ich rola w wielu innych działach językoznawstwa, w tym w gramatyce. Wydawać by się mogło, że tutaj znacznie lepiej odwołać się do kompetencji językowej niż do tekstów. W tych ostatnich bowiem nie znajdziemy raczej niczego ponad to, co jest znane intuicji rodzimego użytkownika języka; co więcej, w tekstach powstałych na użytek normalnej komunikacji międzyludzkiej nie znajdziemy wielu zjawisk językowych akceptowanych przez kompetencję użytkownika języka.

Najprościej na te wątpliwości można odpowiedzieć, że korpus daje nam informacje trojakiego rodzaju. Po pierwsze, jedynie z korpusu dają się zebrać dane statystyczne. Ich wartość jest sprawą dyskusyjną. Znane jest powiedzenie przypisywane Chomsky'emu: „cóż z tego, iż częstsze jest zdanie *Mieszkam w Nowym Jorku* niż *Mieszkam w Dayton, Ohio* – to drugie jest rzadsze, bo więcej ludzi mieszka w Nowym Jorku". Z pewnością sama statystyka nie tłumaczy niczego, dopiero jej wyniki mogą podlegać interpretacji. Ponadto musi ją poprzedzać analiza jakościowa, choćby na tym podstawowym poziomie, jakim jest segmentacja tekstu na słowoformy. Jednak nawet najmniej subtelne dane liczbowe (np. frekwencja danej konstrukcji czy formy fleksyjnej) w jakimś stopniu poszerzają naszą wiedzę o gramatyce. Tym bardziej można to powiedzieć o zaawansowanych technikach, które umożliwiają wykrycie niedostrzegalnych inaczej zależności.

Drugim istotnym obszarem dociekań, gdzie korpus jest niezastąpiony, jest badanie związków gramatyki ze zjawiskami pozagramatycznymi: leksyką, pragmatyką, dyskursem, stylistyką itd.

Trzecim wreszcie obszarem jest badanie konkurencji blisko- lub jednoznacznych form czy konstrukcji. Mogą to być warianty fleksyjne (np. *uczniów* i *uczni*)

albo synonimiczne konstrukcje składniowe (*mówić* + fraza przyimkowa *do kogo* i *mówić* + fraza rzeczownikowa w celowniku). Zwykle rodzimy użytkownik języka nie jest w stanie wskazać warunków, które rządzą wyborem jednej z konstrukcji, mimo że niewątpliwie nieuświadomiona znajomość tych warunków jest częścią jego kompetencji językowej[1].

Niniejszy rozdział, choćby z racji objętości, nie zastąpi podręcznika językoznawstwa korpusowego. Chcę w nim jedynie zasygnalizować możliwe obszary badań gramatycznych z użyciem korpusów, a także pokazać kilka przykładowych zapytań pozwalających poszukiwać danych dotyczących morfologii i składni. W tym zakresie niniejszy rozdział jest rozwinięciem podręcznika użytkownika, zaleca się jego czytanie dopiero po zapoznaniu się ze składnią zapytań. Ograniczam się tutaj do omawiania pracy z korpusem za pomocą wyszukiwarki Poliqarp, ponieważ obsługuje ona bardziej skomplikowane zapytania[2].

Ze względu na to, że niniejszy rozdział jest rozszerzeniem instrukcji użytkownika przykłady będą raczej dość banalne, a pytania pozostawione bez odpowiedzi.

Zacznijmy od tego, że w językoznawstwie korpusowym wyróżnia się trzy podejścia, a mianowicie badania:

1. ilustrowane przez korpus (ang. *corpus illustrated*),
2. oparte na korpusie (ang. *corpus based*),
3. sterowane korpusem (ang. *corpus driven*).

Podejście 1. oznacza tyle, że w argumentacji posługujemy się przykładami pochodzącymi z korpusu zamiast tworzonych ad hoc. Korpus więc w tym wypadku pełni dokładnie taką samą rolę jak informator – rodzimy użytkownik języka. Korpus jednak nie zastąpi nigdy informatora, ponieważ nie da ewidencji negatywnej – z faktu, że jakaś konstrukcja, słowo czy forma nie ma potwierdzenia w korpusie nie wynika, że jest ona nieakceptowana. Badaniami opartymi na korpusie (w sensie 2.) nazywa się takie badania, które mają na celu zweryfikowanie postawionej wcześniej hipotezy. Od pierwszego podejścia różnią się tym, że nie poszukujemy pojedynczej ilustracji postawionej tezy, ale przeszukujemy cały korpus po to, by tę tezę z jednej strony potwierdzić bądź sfalsyfikować, z drugiej, by stwierdzić, czy nie jest ona pozbawiona luk (stawiając hipotezę, nie przewidzieliśmy wszystkich

[1] Zapewne wybór jednej z form leksemu *uczeń* zależy od idiolektu użytkownika języka, stąd wskaże on jedną z form jako akceptowalną. Jeśli jednak chodzi o struktury dla których pozycję otwiera *mówić*, to są one jednakowo akceptowalne dla użytkowników języka. W istocie jednak dają się zaobserwować wyraźne tendencje rządzące wyborem jednej z konstrukcji.

[2] Zalety wyszukiwarki PELCRA ujawniają się z kolei w zastosowaniach leksykologicznych, przede wszystkim ze względu na moduł kolokacyjny, jak i prezentację tzw. profilu historycznego leksemu.

zjawisk). Po trzecie wreszcie, by określić zasięg zjawiska, stwierdzić, czy jest ono typowe, czy marginalne[3].

Z kolei podejście 3. charakteryzuje się brakiem jakichkolwiek hipotez stawianych przed przystąpieniem do badań. Jego istotą ma być taksonomiczne budowanie gramatyki na podstawie danych empirycznych pozyskiwanych z korpusu. W konsekwencji oczekuje się od tego rodzaju podejścia opisu neutralnego pod względem teoretycznym.

Istotne jest rozróżnienie między podejściem 1. z jednej strony a 2. i 3. z drugiej. Granica między 2. i 3. nie jest bardzo ostra, tym bardziej, że wydaje się, iż ortodoksyjnie prowadzone badania „sterowane korpusem" są niemożliwe, nie sposób bowiem zaczynać przeszukiwania korpusu bez postawienia jakiejś (choćby roboczej) hipotezy, która często jest doprecyzowywana w trakcie obserwacji korpusu. W praktyce więc zazwyczaj badania korpusowe są po części weryfikacją przyjętych hipotez, po części badaniem całkowicie indukcyjnym.

Jak widać, nie ma czegoś takiego jak ściśle określona metodologia językoznawstwa korpusowego. Przeciwnie, korpus jest jedynie narzędziem (a ściślej źródłem pewnego rodzaju danych) i różni językoznawcy czynią zeń bardzo różny użytek. Z drugiej strony charakter tych danych sugeruje (bo przecież nie narzuca) pewien styl badań. Będzie to więc analiza raczej ilościowa niż tylko jakościowa, bardziej poszukiwanie tendencji niż reguł. Wreszcie będzie to rozpatrywanie równocześnie formy i funkcji bytów gramatycznych.

Dla bardzo wielu pytań o gramatykę punktem wyjścia będzie poszukiwanie konkretnego leksemu. Dość powiedzieć, że do odnalezienia przykładów zdań podrzędnych wystarczy zapytanie `[base = "że|iż|żeby"]`. Oczywiście nie da to nam wszystkich wystąpień zdań podrzędnych choćby dlatego, że to zapytanie nie zwróci nam zdań z *kiedy*. Z kolei zapytanie `[base = "że|iż|żeby|kiedy"]` zwróci sporą liczbę zdań, które nie są tym, o co pytamy np.:

(17.1) *Że też ludzi interesują jeszcze takie głupstwa – mruknął (…)* (Włodzimierz Kowalewski *Światło i lęk. Opowiadania starej daty*).

(17.2) *A kiedy przyjdziesz?* (Hanna Samson *Pułapka na motyla*).

I tu dochodzimy do dylematu, jaki towarzyszy formułowaniu każdego zapytania w korpusie: czy wolimy zwiększać dokładność, czy pełność? Innymi słowy, czy wolimy jedynie rzeczywiste przykłady kosztem przeoczenia wielu wystąpień badanego zjawiska (czyli osiągnąć wysoką dokładność), czy przeciwnie, chcemy wszystkich wystąpień kosztem tego, że prócz tych, które nas

[3] U wielu badaczy badania typu 2. i 3. nie są wyróżniane i nazywa się je badaniami opartymi na korpusie (*corpus based*), por. Lewandowska-Tomaszczyk 2004.

interesują, będą również takie, o które nam wcale nie chodziło (a więc chcemy osiągnąć wysoką pełność).

Co gorsza, w praktyce nigdy nie otrzymamy ani pełnej dokładności (por. przykład (17.1)), ani też całkowitej pełności. Przy tym najczęściej po osiągnięciu pewnego pułapu niewielki wzrost pełności powoduje radykalny spadek dokładności. Z czysto praktycznego punktu widzenia jest to jednak ważna decyzja: jeśli szukamy pojedynczego przykładu, istotniejsza jest dokładność. Jeśli chcemy całościowo przebadać jakieś zjawisko w korpusie, to istotniejsza jest pełność. W tym wypadku jednak często konieczna jest „ręczna" selekcja otrzymanego materiału. Należy więc w takim wypadku podjąć decyzję, czy uznajemy, że można wyciągnąć wnioski na podstawie niepełnego zestawu wystąpień badanego zjawiska i wtedy postarać się zwiększyć dokładność kosztem pełności, czy też przeciwnie. Wprawdzie dobrze skonstruowane zapytanie[4] często pozwala wydobyć bardzo znaczącą część (powyżej 90%) wystąpień szukanego zjawiska, ale zawsze pozostaje pytanie, czy te nietypowe wystąpienia, niepasujące do wzorca wyszukiwania, nie różnią się w istotny sposób od typowych. Niezależnie od decyzji zawsze powinno się oszacować tak pełność, jak i dokładność.

Wróćmy do wyszukiwania: w dociekaniach gramatycznych najczęściej konieczne jest odwołanie się do znakowania morfosyntaktycznego. Przykładowo jak powszechnie wiadomo wyraz *reżyser* ma dwie formy mianownika liczby mnogiej: *reżyserzy* i *reżyserowie*. Do zbadania występowania tych form wystarczy zapytanie `[orth = "reżyserzy|reżyserowie"]`. Wariantywność tego typu jest jednak charakterystyczna dla większej liczby rzeczowników męskoosobowych zakończonych na *-er* lub *-or*, np. *profesor*, *redaktor*. Potrzebne jest nam zapytanie, które zwróci wszystkie takie słowa, nie tylko ze względu na ekonomię działania (chcemy uzyskać odpowiedź za pomocą jednego zapytania), ale także dlatego, że chcemy się równocześnie dowiedzieć, które rzeczowniki wykazują w tekstach takie wahania. Oczywiście można listę słów stworzyć a priori, ale nie wiadomo, czy będzie ona pełna[5]. Rozpatrzmy na początek następującą możliwość: pytamy o zakończenia mianownika liczby mnogiej o postaci *-rzy* oraz *-owie*, posługując się zapytaniem `[orth = ".*rzy|.*owie" & cas = nom & nmb = pl]`. To zapytanie oczywiście zwróci nam *którzy* czy *dobrzy*. Tymczasem pytanie, które postawiliśmy na początku dotyczyło wariantywności końcówek mianownika liczby mnogiej rzeczowników zakończonych na *-er* i *-or*. Należy więc zapytać o wystąpienia spełniające którąś

[4] Oczywiście nie zależy to jedynie od konstrukcji zapytania. Tam, gdzie mamy do czynienia z homografią lub wieloznacznością, nie da się skonstruować zapytania, które oddzieli dwa różne zjawiska językowe. Niech będzie przykładem choćby para zdań: *Janek się myje w ciepłej wodzie* i *Taki talerz się myje w wodzie z ludwikiem*.

[5] Warto zwrócić uwagę, że posłużenie się gotową listą bardziej wpisuje się w nurt językoznawstwa opartego na korpusie, natomiast drugie podejście – językoznawstwa sterowanego korpusem.

z par warunków: forma podstawowa zakończona na *-er* i *-or* i wyraz tekstowy zakończony na *-rzy* lub *-owie* (`[base = ".*er|.*or" & orth = ".*owie|.*rzy"]`) bądź forma podstawowa zakończona na *-er* i *-or* i wykładnik mianownika liczby mnogiej (`[base = ".*er|.*or" & cas = nom & nmb = pl]`).

W wypadku badania wariantywności form fleksyjnych trzeba jednak pamiętać o tym, że nie zawsze należy ufać znakowaniu morfosyntaktycznemu. To ostatnie bowiem dokonywane jest, jak wiadomo, automatycznie; automat działa na podstawie reguł, a te z kolei są tworzone na podstawie gramatyki normatywnej. Nie oznacza to skrajnego puryzmu, jednak automat rozpoznaje jedynie te formy, które przewiduje gramatyka. Tak więc np. zapytanie `[base = "palec" & cas = gen & nmb = "plur"]` zwróci nam tylko formy *palców*, mimo że forma *palcy* jest reprezentowane w korpusie[6].

Podobnie możemy poszukiwać w korpusie żeńskich derywatów od męsko-osobowych nazw zawodów.

Przejdźmy teraz do zagadnień składniowych, a więc do pytań przekraczających granice słowa. Nie omawiam tutaj wykorzystania anotacji składniowej, temu zagadnieniu jest bowiem poświęcony rozdz. 8. Skoncentruję się raczej na wyszukiwaniu konstrukcji, które nie są wyróżnione za pomocą wzmiankowanej anotacji.

Pamiętajmy, że w istocie nie możemy pytać o konstrukcje, a jedynie o określone przez nas sekwencje znaczników. Mogą one zwrócić rzeczywiste wystąpienia poszukiwanej konstrukcji, ale najczęściej zwrócą też ciągi, które tej konstrukcji nie reprezentują.

Zacznijmy od tego, że w pewnych wypadkach pytamy raczej o ciąg leksemów niż form gramatycznych. Jeśli interesuje nas dystrybucja dwu realizacji argumentu czasownika *opowiadać*, tzn. z elementem prosentencjalnym, jak w zdaniu:

(17.3) *(…) opowiada o tym, że człowiek jest urodziwy w swoich pomyłkach* (JOC *"Łoże" i gra z widzem*, „Dziennik Polski")

i bez tego elementu

(17.4) *Opowiadają, że pomysł pojawił się znacznie wcześniej* (tw *Wikarego nie oddali*, „Sztafeta")

wystarczą zapytania nieodwołujące się do oznaczeń gramatycznych:

(17.5) `[base = opowiadać][base = o][base = to][][base = że]`

(17.6) `[base = opowiadać][][base = że]`[7]

[6] Zainteresowanych zagadnieniem wariantywności we fleksji odsyłam do książki Hebal-Jezierskiej (2008).

[7] Uważny czytelnik zapewne zauważy, że w szukanych ciągach jest jeden segment, na który nie nałożono żadnego ograniczenia. Jest to miejsce na przecinek, segment ten mógłby oczywiście mieć postać `[base = "[,]"]`, jednak w tej pozycji, w prawidłowo zbudowanym zdaniu nie pojawi się nic innego, stąd dla ułatwienia można zastosować proponowany przeze mnie zapis. Notabene zapytanie `[base = opowiadać][base = że]` również zwróci pewną liczbę kontekstów obarczonych błędem ortograficznym.

Zwróćmy uwagę na to, że obie konstrukcje są równoważne, powstaje więc pytanie, czy język toleruje istnienie tego rodzaju synonimii składniowej, czy też obie konstrukcje wykazują jakieś różnice.

Kolejny przykład to pytania o możliwą długość ciągu złożonego z samych rzeczowników lub przymiotników z dopełniaczu. Jakkolwiek takie pytanie samo w sobie jest niezbyt poważne, to odpowiedź na nie ma istotniejsze znaczenie, a mianowicie pośrednio mówi nam, jak bardzo można rozbudowywać frazę nominalną. Okazuje się, że korpus notuje przykłady z 9 takimi segmentami

(17.7) *(…) zarządu wojskowego oddziału kozackiego ministerstwa Rzeszy okupowanych ziem wschodnich* (*Siemion Krasnow*, Wikipedia)

które zwraca zapytanie `[cas = gen & !pos = prep]{9}`. Oczywiście frazy taki są rzadkie; zmieniając cyfrę w tym zapytaniu, możemy zbadać zależność liczby tego rodzaju fraz od stopnia ich rozbudowania.

Rozważmy następujący problem: w języku polskim tzw. czas przyszły złożony ma dwie postaci: forma czasu przyszłego od czasownika *być* + bezokolicznik bądź pseudoimiesłów. Możemy postawić hipotezę, że jeśli podmiotem jest rzeczownik rodzaju żeńskiego, wystąpi raczej pierwsza z tych form[8]. Hipoteza ta ma oczywiste uzasadnienie: ma ona z zasady o jedną sylabę więcej por. *będę pisała : będę pisać*. Spróbujmy doprecyzować naszą hipotezę: tendencja ta może się ugruntować tylko wtedy, gdy mówiący wyraźnie częściej używa formy żeńskiej, co ma miejsce, gdy jest kobietą, ograniczmy więc przeszukiwanie do 1. osoby liczby pojedynczej. Ponieważ liczba sylab w obu wariantach dla rodzaju męskiego jest identyczna *(będę pisał : będę pisać)*, opisywany mechanizm nie zadziała w wypadku, gdy mówiącym jest mężczyzna. Nie działa on też w liczbie mnogiej. Spróbujmy więc zweryfikować tę hipotezę. Jak się wydaje najprostszy sposób to porównać liczbę wystąpień jednej i drugiej formy w korpusie. Można to zrobić za pomocą zapytań:

(17.8)
```
[pos = bedzie & nmb = sg & per = pri] [pos = praet & gnd =
m1] | [pos = praet & gnd = m1] [pos = bedzie & nmb = sg &
per = pri],
```

co należy interpretować jako: wyszukaj wystąpienia sekwencji segmentów a) czas przyszły czasownika *być* w liczbie pojedynczej i pierwszej osobie oraz b) pseudoimiesłów rodzaju męskoosobowego lub sekwencji tych samych segmentów, ale w odwrotnej kolejności. Następnie zmieniamy odpowiednio w zapytaniu rodzaj na żeński. W ten sposób wyszukujemy wystąpienia form typu *będę pisał, będę widział* oraz *będę pisała, będę widziała*. Jeżeli okaże się, że liczba opisanych form jest znacząco wyższa dla rodzaju męskoosobowego niż żeńskiego, to możemy próbować wyciągać z tego wniosek, że hipoteza postawiona na początku tego wywodu jest prawdziwa.

[8] Zagadnienie to badał Dunaj (1987).

Niezależnie od rzeczywistych wartości należy uznać wyniki za niewiarygodne. Nawet jeśli się okaże, że faktycznie liczba kontekstów zwróconych przez pierwsze z zapytań będzie wyższa niż przez drugie, to bardzo możliwe, że wynika to stąd, iż korpus notuje więcej wypowiedzi mężczyzn niż kobiet. Co gorsza, w komponencie pisanym nie badamy przecież wypowiedzi kobiet, ale najczęściej fikcyjne wypowiedzi pisane przez autora dowolnej płci, które przypisuje on bohaterom – kobietom. Oczywiście dotyczy to również rzeczywistych wypowiedzi kobiet (np. w wywiadach), które zostały poddane obróbce redakcyjnej.

Wiarygodnych danych dostarcza korpus mówiony. W tym celu powinniśmy przeszukać go za pomocą zapytania:

(17.9) `[pos = bedzie & nmb = sg & per = pri] [pos = inf] | [pos = inf][pos = bedzie & nmb = sg & per = pri]`

w wypowiedziach mężczyzn i w wypowiedziach kobiet, korzystając z ograniczenia wyszukiwania za pomocą metadanych. Dopiero porównanie tych czterech liczb, tzn. liczby: a) konstrukcji z pseudoimiesłowem użytych przez kobiety, b) konstrukcji z bezokolicznikiem użytych przez kobiety, c) konstrukcji z pseudoimiesłowem użytych przez mężczyzn, d) konstrukcji z bezokolicznikiem użytych przez mężczyzn, pozwoli zweryfikować naszą hipotezę. Pamiętajmy wszakże, że pytamy o zapewne najbardziej typową strukturę, tzn. taką, w której słowo pomocnicze i czasownik występują obok siebie. Jest jednak niemała liczba przykładów, gdzie jedno od drugiego oddziela kilka segmentów, np.[9]:

(17.10) *Nie będę w związku z tym tego powtarzać* (Sprawozdanie stenograficzne z obrad Sejmu RP z dnia 31.08.1995, 2 kadencja, 58 posiedzenie, 2 dzień)

Pozostaje pytanie czy można wyniki uzyskane za pomocą takiego najbardziej typowego zapytania przenieść na cały korpus, czy przeciwnie – gdy słowo pomocnicze jest oddalone od czasownika, działają inne mechanizmy. Jak się wydaje, obie sytuacje są możliwe.

Rozważmy teraz inny, bardziej skomplikowany wzorzec wyszukiwania, który ma zwrócić czasowniki w stronie biernej. Na potrzeby zapytania możemy stronę bierną zdefiniować jako sekwencję imiesłowu biernego i czasownika posiłkowego *być* lub *zostać* (w dowolnej kolejności). Odpowiada to zapytaniu:

(17.11) `[pos = ppas][base = "być|zostać"] | [base = "być|zostać"] [pos = ppas].`

Zapytanie to zwróci nam również zdania typu:

(17.12) *Widziałaś jaki był zmoknięty?* (Maria Krüger *Witaj Karolciu*),

które nie są przykładami strony biernej. Tego rodzaju przykłady najprościej odfiltrować ręcznie. Jeśli jednak zapytanie będzie powtarzane wielokrotnie, choćby

[9] Przykład ten został wyszukany zapytaniem `[pos = bedzie] [!pos = verb]{5} [pos = inf] within s`. Oczywiście zapytanie to dało bardzo dużo zdań, które nie były przykładami łącznika i orzecznika oddzielonego od siebie pięcioma segmentami.

z niewielkimi zmianami, warto rozważyć stworzenie listy tych czasowników, które wchodzą w grę, bądź listy negatywnej (tzw. stoplisty). W tym drugim wypadku zapytanie miałoby postać:

(17.13) `[pos = ppas & !base = "ubrać|zmoknąć|itp."][base = "być|zostać"] | [base = "być|zostać"] [pos = ppas & !base = "ubrać|zmoknąć|itp."].`

Jakkolwiek długość zapytania jest ograniczona, Poliqarp powinien przyjąć zapytanie z przynajmniej setką wyrazów ze stoplisty[10].

Z drugiej strony nie zwróci nam ono bardzo wielu przykładów, w których czasownik posiłkowy jest oddzielony pewną liczbą segmentów, choćby końcówką fleksyjną[11], jak np.:

(17.14) *(…) przez nikogo nie byłem ścigany* (Michał Bielecki (Stanisław Bieniasz) *Dziewczyna z Banku Prowincjonalnego S. A.*)

Powinniśmy więc zmodyfikować zapytanie tak, by dopuścić pewną liczbę segmentów między czasownikiem posiłkowym a imiesłowem, powiedzmy do nie więcej niż dwu (dla czytelności wywodu pomijam listę czasowników):

(17.15) `[pos = ppas] [] {,2} [base = "być|zostać"] | [base = "być|zostać"] [] {,2} [pos = ppas].`

Zapytanie to zwróci nam zdanie:

(17.16) *(…) ofiara nie została jednak nawet zauważona* (Wojciech Jagielski *Modlitwa o deszcz*),

ale także:

(17.17) *(…) na których czele będzie kroczył, niezdeprawowany (…)* (Wojciech Jagielski *Modlitwa o deszcz*),

(17.18) *Na pochyłym brukowanym rynku odnaleźliśmy przystanek autobusu.* (Andrzej Stasiuk *Biały kruk*).

Aby odfiltrować powyższe zdania, należy zmodyfikować zapytanie, dodając ograniczenie, że żaden z tych dwu segmentów wstawionych między imiesłów a słowo pomocnicze nie może być czasownikiem. Z kolei, żeby wyeliminować drugie ze zdań, należy nałożyć ograniczenie, że czasownik *być* nie może być aglutynatem (w tym wypadku wyraz *być* jest realizowany jako *-śmy*).

[10] Naturalnie wpisywanie tak długich zapytań bezpośrednio do okienka w interfejsie użytkownika jest skrajnie niepraktyczne. Powinno się je raczej formułować w edytorze tekstu, a następnie kopiować. Pomijając to, że edycja jest łatwiejsza, można tego rodzaju zapytań używać wielokrotnie. Zabezpiecza to też przynajmniej częściowo przed drobnymi błędami literowymi, które powodują, że zapytanie zwraca wyniki fałszywe z punktu widzenia użytkownika.

[11] W tym miejscu z naciskiem chciałbym przypomnieć o tym, że ruchoma końcówka fleksyjna ma status osobnego segmentu, o czym bardzo łatwo zapomnieć, tworząc zapytanie, w którym jakiś segment ma następować bezpośrednio po czasowniku.

Omawiany wzorzec wyszukiwania można by dalej ulepszać. Zasadniczo daje się to robić tylko eksperymentalnie, tzn. za każdym razem oceniając, na ile wzrasta dokładność i pełność zwracanych przez wyszukiwarkę kontekstów. W końcu należy się zatrzymać w miejscu, gdzie dramatycznie spada dokładność, a pełność poprawia się już nieznacznie.

Skoro mowa o stronie biernej, to zauważmy, że powszechnie przyjmuje się, że uzasadnieniem dla strony biernej z jawnie wyrażonym agensem jest pragmatyka, a ściślej potrzeba tematyzacji patiensa. Taką tezę można zasadniczo zweryfikować tylko przy użyciu korpusu. Zwracam uwagę, że o ile w przykładach dotąd przytaczanych wystarczyła obserwacja konkordancji, tutaj należy korzystać z podglądu szerszego kontekstu[12].

Jak wcześniej wspomniałem, jedno z tych pól badań językoznawczych, w których korpus jest niezastąpiony, to zagadnienie zróżnicowania stylistycznego języka. Wprawdzie świadomość tego zróżnicowania istniała wcześniej, jednak dopiero elektroniczne korpusy pozwoliły głębiej wejść w omawiane zagadnienie. Wyszukiwarki NKJP umożliwiają stworzenie podkorpusów definiowanych przez użytkownika. Jest to możliwe dzięki opcji ograniczenia wyszukiwania za pomocą metadanych. Zapewne największe zastosowanie znajdzie podział na typ tekstu. Oczywiście przyjęta przez nas typologia nie jest jedyną możliwą. Z pewnym nakładem pracy użytkownik może definiować własne podkorpusy tworząc listę tekstów przynależnych do określonej przez siebie kategorii. Można więc np. podzielić literaturę piękną na „popularną" i „kierowaną do wąskiego grona odbiorców" (jakkolwiek byłyby te kategorie definiowane). Interesujące może być zarówno całościowe badanie różnic pomiędzy rejestrami i odmianami języka jak i odmienność poszczególnych zjawisk językowych w różnych rejestrach.

Dodajmy, że jakkolwiek NKJP jest korpusem synchronicznym, najstarsze teksty, które wchodzą w skład części niezrównoważonej liczą ponad 100 lat. Może więc pełnić funkcję korpusu diachronicznego, w którym dokonywane będą porównania tekstów dziewiętnastowiecznych ze współczesnymi.

[12] W pracy Górski 2008 szerzej omawiam metody wypracowane przez Givona (1983), zwane trwałością tematu (ang. *topic persistence*) i odległością referenta (ang. *referential distance*).

18

NKJP w warsztacie tłumacza

Piotr Pęzik

18.1. Rola korpusów referencyjnych

Narodowy Korpus Języka Polskiego (NKJP) to korpus referencyjny. Choć kuszące wydaje się zdefiniowanie korpusu referencyjnego jako korpusu reprezentującego *ogólny rejestr języka*, to bezpieczniej będzie chyba powiedzieć, że ma on w założeniu reprezentować użycie bardzo różnych rejestrów i typów funkcjonalnych języka. Warto o tej podstawowej funkcji NKJP pamiętać, próbując określić jego rolę w warsztacie tłumacza. Współcześni tłumacze posługują się bowiem bardzo różnymi narzędziami i zasobami, których korpus referencyjny często zastąpić nie może. Na przykład pracujący w projektach zespołowych tłumacze na co dzień korzystają z wyspecjalizowanego oprogramowania umożliwiającego tworzenie i aktualizację pamięci tłumaczeniowej oraz baz terminologicznych, które podnoszą stopień spójności językowej oraz merytorycznej tłumaczeń. Przykładem szerokiego wykorzystania wspomagających tłumaczenie technologii są narzędzia do lokalizacji oprogramowania, gier komputerowych, serwisów internetowych jak również innych treści elektronicznych.

Postęp w dziedzinie technologii informacyjnych zmienił zresztą nie tylko metody weryfikacji terminologicznej, czy też szerzej mówiąc lingwistycznej tłumaczonych tekstów. W dobie Internetu tłumacze zupełnie inaczej niż w przeszłości zaspokajają swoje potrzeby informacyjne. Zgodnie z łacińską sentencją *Rem tene, verba sequentur*, jakość tłumaczenia zależy przede wszystkim od zrozumienia tekstu źródłowego, a dopiero w dalszej kolejności od kompetencji językowych tłumacza. Swoje doraźne potrzeby informacyjne tłumacze zaspokajają dziś w pierwszym odruchu za pomocą popularnej wyszukiwarki internetowej albo tworzonych społecznościowym wysiłkiem encyklopedii, takich jak Wikipedia. Dla tłumaczy języków szeroko w niej reprezentowanych (m. in. języka polskiego) ta

wielojęzyczna encyklopedia staje się często źródłem informacji o ekwiwalentach terminologicznych. Wikipedia przewyższa bowiem tradycyjne słowniki i leksykony specjalistyczne pod względem łatwości dostępu i aktualności informacji oraz jednoznaczności sugerowanej ekwiwalencji. Wynika to głównie z faktu, że połączone hipertekstem różne wersje językowe artykułów encyklopedycznych umożliwiają dużo lepszą weryfikację merytoryczną odpowiedników terminologicznych niż słownik, który podaje zazwyczaj jedynie pary ekwiwalentów.

NKJP nie jest bazą terminologiczną, encyklopedią ani zbiorem tekstów specjalistycznych z konkretnej dziedziny. Z pewnością nie może też konkurować z zasobami Internetu jako źródło wiedzy o tematyce tłumaczonych tekstów. Korpusowi referencyjnemu nie można też przypisywać roli korpusu równoległego, który umożliwia bezpośrednie porównywanie zrównoleglonych jednostek tłumaczeniowych w różnych językach. Jaką więc funkcję może pełnić korpus referencyjny w warsztacie tłumacza? Głównie taką, jaka wynika z jego podstawowej funkcji jako dużego, anotowanego bibliograficznie, strukturalnie i lingwistycznie, zrównoważonego zbioru tekstów pisanych i mówionych. Odpowiednio duży korpus referencyjny stanowi niezastąpione źródło informacji o użyciu, rejestrze, konotacjach oraz łączliwości semantycznej i frazeologicznej wyrazów, zwrotów, form i konstrukcji gramatycznych. Choć zapewnienie ekwiwalencji tłumaczenia na tych poziomach zależy głównie od intuicji i nabytych umiejętności tłumacza, to jednak korpus referencyjny może być dużo skuteczniejszym narzędziem oceny równoważności frazeologicznej tłumaczeń niż jakikolwiek słownik. W tym rozdziale skupiam się w głównej mierze na przydatności korpusu referencyjnego w weryfikacji poprawności i ekwiwalencji *frazeologicznej* oraz *leksykalno-gramatycznej* tłumaczeń.

18.2. Ekwiwalencja frazeologiczna

Kryteria i poziomy równoważności tłumaczonych tekstów należą do podstawowych zagadnień traduktologii, których śladów możemy się doszukiwać w starożytności. Już w *Liście do Pizonów* Horacego znajdujemy przestrogę przed „ciasną imitacją" dosłownego tłumaczenia. Współcześni teoretycy translacji, poza rozróżnieniem między ekwiwalencją formalną i dynamiczną (Nida i Taber 1969), definiują też poszczególne poziomy funkcjonalne równoważności między tłumaczonymi tekstami, na przykład ekwiwalencję leksykalną, gramatyczną, tekstualną, czy też pragmatyczną (Baker 1992). Oczywiście w praktyce tłumaczenia różne poziomy ekwiwalencji mają charakter doraźny i nie powinny sugerować istnienia stałych dla każdego tekstu strukturalnych jednostek tłumaczeniowych. Niemniej jednak bez przyjęcia w danym kontekście określonego poziomu równoważności trudna byłaby jakakolwiek szczegółowa analiza porównawcza oryginału i przekładu.

O ekwiwalencji frazeologicznej można mówić pośrednio jako o rodzaju ekwiwalencji leksykalnej, ale dużo bardziej uzasadnione wydaje się wyróżnienie frazeologii jako osobnego wymiaru ekwiwalencji. Niemal każdy tekst składa się bowiem w dużej mierze z bardziej lub mniej utrwalonych połączeń frazeologicznych, z których wiele nie znalazło się w żadnym słowniku. Jednym z zadań tłumacza jest zachowanie równowagi między oryginałem a przekładem w warstwie kolokacji, terminów, rzeczowników wielowyrazowych czy też idiomów. Trudność tego zadania widać zwłaszcza w tłumaczeniach studentów translatoryki, którzy działając pod przemożnym wpływem tekstu oryginalnego, kompozycyjnie oddają oryginalne frazeologizmy jako złożenia wyrazów, które w języku przekładu nie mają statusu frazeologicznego. Taką nierówność frazeologiczną w tłumaczeniu bardzo trudno jest wykazać bez odwołania się do reprezentatywnych zbiorów danych językowych jakimi są korpusy referencyjne. Dużo łatwiej udowodnić brak równoważności na poziomie leksemu, czy nawet niektóre błędy rzeczowe, niż brak odpowiednika frazeologicznego w tłumaczeniu. Z jednej strony tłumacz może nie rozpoznać połączenia wyrazów w tekście oryginalnym jako utrwalonego frazeologizmu, ale chyba jeszcze częściej, działający pod presją czasu tłumacz nietrafnie, lub zbyt kompozycyjnie oddaje utrwalony frazeologizm, mimo iż istnieje jego dużo lepszy odpowiednik. Wynika to w dużej mierze z kognitywnego mechanizmu dostępu do wiedzy frazeologicznej użytkowników danego języka. Autor tego rozdziału wielokrotnie przeprowadzał eksperyment, w ramach którego grupa rodzimych użytkowników polszczyzny proszona była o wypisanie kolokacji przymiotnikowych danego wyrazu, np. rzeczownika *nos*. Przeciętny uczestnik eksperymentu był w stanie podać nie więcej niż 10 stosunkowo mocno utrwalonych kolokacji w ciągu kilku minut, podczas gdy proste zapytanie wpisane do Kolokatora NKJP zwraca kilkadziesiąt statystycznie istotnych[1] kolokacji tego rzeczownika[2]. Można zatem wnioskować, że bierna znajomość frazeologii nie zawsze przekłada się bezpośrednio na wybory frazeologiczne tłumacza, który pracuje w warunkach sprzyjających interferencji języków oryginału i przekładu. Zilustrujmy to autentycznym przykładem tłumaczenia artykułu poświęconego kontrowersjom naukowym i politycznym wokół przyczyn globalnego ocieplenia („The Economist" 2009)[3]. Bezpośrednio pod tytułem artykułu (*A heated debate*), pojawia się podtytuł, który nastręczył niektórym studentom translatoryki tłumaczącym ten tekst trudności z zachowaniem równoważności frazeologicznej między oryginałem (18.1) a przekładem (18.2).

[1] Zob. rozdz. 14.

[2] Zob. `http://nkjp.uni.lodz.pl/?q=5udfcdw`.

[3] Tekst ten był dostępny w czasie pisania artykułu w internetowym wydaniu tygodnika, pod adresem: `http://www.economist.com/node/14966227?story_id=14966227`.

(18.1) O: Why political orthodoxy must not silence *scientific argument*.
(18.2) T: Dlaczego ortodoksja polityczna nie może uciszyć *kłótni naukowej*.

Zwróćmy uwagę na dwa, w różnym stopniu utrwalone frazeologizmy angielskie pojawiające się w tym krótkim podtytule. Po pierwsze, pojawia się tutaj kolokacja *political orthodoxy*, która została oddana jako *ortodoksja polityczna*. Mimo że w angielszczyźnie ta kolokacja jest silniej reprezentowana, niż na przykład w NKJP, to jednak obie wersje językowe tej frazy wydają się mieć pewien status frazeologiczny. Typowe kolokaty rzeczownika *ortodoksja* to przymiotniki określające przynależność religijną lub ideologiczną (tab. 18.1)[4]. Niektóre z tych połączeń

Tabela 18.1. Kolokacje przymiotnikowe rzeczownika *ortodoksja* w NKJP

Nr	Kolokacja	Przykład	Źródło
1.	katolicka	Mnóstwo jest obaw, że Benedykt XVI nadaje nową siłą *katolickiej ortodoksji*. Że wraca do przedsoborowych wzorców i metod	Blog Jarosława Dudy
2.	nicejska	W tych doktrynalnych walkach kluczową rolę odegrał św. Atanazy, obrońca *ortodoksji nicejskiej*.	*Dzieje chrześcijaństwa w zarysie*
3.	marksistowska	Chociaż do śmierci odgrywał rolę strażnika *marksistowskiej ortodoksji*, przyczynił się do powstania rewizjonizmu.	„Gazeta Wyborcza"
4.	religijna	(...) dzisiaj chasydyzm jest synonimem *ortodoksji religijnej*, życia według precyzyjnie przestrzeganych praw (...)	„Gazeta Wyborcza"
5.	polityczna	(...) Mam nadzieję, że w tej walce sacrum i profanum nie zwycięży niemoralna i *polityczna ortodoksja*.	„Dziennik Polski"

można wręcz uznać za terminy (2), inne są luźniej utrwalonymi kolokacjami. Określenie *ortodoksja polityczna* wydaje się w tym kontekście dobrym odpowiednikiem angielskiej frazy *political orthodoxy*, co potwierdzają przykłady angielskich kolokacji rzeczownika *orthodoxy*, zaczerpnięte z Brytyjskiego Korpusu Narodowego (BNC), ukazane w tab. 18.2[5].

Bardziej problematyczne natomiast wydaje się tłumaczenie występującej w angielskim oryginale frazy *scientific argument*. Po pierwsze, angielski rzeczownik *argument* można zinterpretować jako *spór*, *kłótnię* (tab. 18.3, p. 1, 2, 3) lub też

[4] Zob. `http://nkjp.uni.lodz.pl/?q=6zkrusn`.
[5] Zob. `http://nkjp.uni.lodz.pl/?q=667c8ju`.

Tabela 18.2. Kolokacje przymiotnikowe rzeczownika *orthodoxy* w BNC

Nr	Kolokacja	Przykład	Źródło
1.	Christian	The official imposition of *Christian orthodoxy* at the end of the century reinforced the tendency to merge church and empire (...)	BNC, plik ADC
2.	communist	the waves of perestroika lap at one of the last fortresses of *communist orthodoxy*.	BNC, plik A3T
3.	Catholic	they found a guardian of *Catholic orthodoxy* saying that he hoped and expected to meet atheists in heaven.	BNC, plik A68
4.	economic	*Economic orthodoxy* over the past decade has rightly stressed the role of the private sector in development.	BNC, plik A68
5.	political	The inculcation of *political orthodoxy* and instruction of a more coercive nature was left strictly in the hands of the Party.	BNC, plik A64

po prostu *argument* (tab. 18.3, p. 4, 5, 6). To ostatnie znaczenie byłoby ze względów gramatycznych mniej prawdopodobne (*argument* w tym znaczeniu jest zazwyczaj policzalny i w liczbie pojedynczej występuje z przedimkiem), gdyby nie typowe w angielskich tytułach prasowych pomijanie przedimków. Co ciekawe, jeżeli przyjmiemy, że fraza *scientific argument* jest kolokacją, to mamy do czynienia z wyjątkiem od stosowanej w systemach automatycznego ujednoznaczniania słów reguły jednoznaczności kolokacji (Yarowsky 1993). Jednak dokonana przez tłumacza interpretacja znaczenia tego wyrazu wydaje się słuszna w kontekście całego artykułu, który mówi raczej o trwających od wielu lat sporach, niż o konkretnym argumencie w kwestii globalnego ocieplenia. Jednakże użyta do oddania tego znaczenia fraza *kłótnia naukowa* uderza jako zbyt kompozycyjne w porównaniu z oryginałem złożenie wyrazów, które narusza równoważność frazeologiczną między oryginałem a przekładem. *Kłótnia naukowa*, a w zasadzie *naukowa kłótnia* występuje w całym zbiorze danych NKJP tylko jeden raz. Przymiotnika *naukowa* nie znajdziemy wśród typowych kolokatów przymiotnikowych rzeczownika *kłótnia* (takich jak *pijacka*, *wewnątrzkoalicyjna*, *ciągła*, *małżeńska*, *gorsząca*, *zażarta*, *sąsiedzka*, *gwałtowna*, *bezproduktywna* itp.)[6]. Analizując za pomocą korpusu referencyjnego języka oryginału oraz języka przekładu użycie dwóch fraz, które

[6] Warto zauważyć, że typowe kolokaty *kłótni* nacechowane są negatywnie; domyślnie *kłótnia* nie ma konstruktywnych konotacji. Zob. `http://nkjp.uni.lodz.pl/?q=6xuxkxc`.

w zamierzeniu tłumacza miały być równoważne, można dojść do wniosku, że mamy do czynienia z brakiem ekwiwalencji frazeologicznej. Kolokacja występująca w oryginale nie ma dobrego odpowiednika w przekładzie, ponieważ została oddana jako fraza o znikomej lub wręcz żadnej łączliwości frazeologicznej, w dodatku wywołującej inne od oryginalnych prozodie semantyczne. Wydaje się, że w NKJP można znaleźć dużo lepsze odpowiedniki *scientific argument*. Setki wystąpień fraz *debata naukowa*, *spór naukowy* czy też *dyskusja naukowa* potwierdzają ich utrwalony status frazeologiczny, którego brakuje *kłótni naukowej*[7].

Tabela 18.3. *Scientific argument* w BNC

Nr	Przykład	Źródło
1.	As you may know, there has been much *scientific argument* about the ways such traits appear.	BNC, plik AN8
2.	The other *scientific argument* raging at the conference involved the start of bipedality.	BNC, plik B7G
3.	and persuasive participant at meetings and was hard to defeat in *scientific argument* — not least because he often interpreted data more correctly	BNC, plik EAK
4.	make swifter judgments than a lay jury on the weight and validity of *scientific arguments*.	BNC, plik EC7
5.	which made them superior to coloured people, and all kind of bogus *scientific arguments* followed from that.	BNC, plik KRF
6.	What roughly he says is that he is committed to the following *scientific arguments*, arguments based upon scientific principles, regardless of any pain	BNC, plik KS3

Tabela 18.4. *Naukowa kłótnia* w NKJP

Nr	Przykład	Źródło
1.	Od kilku lat trwają badania, dyskusje i *naukowe kłótnie* na temat tego, jak zwalczyć sinice - dziwi się Joachim Gorus (...)	„Nowa Trybuna Opolska" (NKJP)

18.3. Poprawność leksykalno-gramatyczna przekładu

W lingwistyce korpusowej używa się czasem terminu *leksykogramatyka* (ang. *lexicogrammar*) do opisania ogółu relacji między jednostkami leksykalnymi a klasa-

[7] Zob. `http://nkjp.uni.lodz.pl/?q=5tvys7v`.

mi gramatycznymi (Halliday 2004). Wiele relacji leksykalno-gramatycznych ma charakter tendencji a nie reguły i trudno w pełni wytłumaczyć, lub nawet zidentyfikować bez odniesienia do reprezentatywnych zbiorów danych językowych, jakimi są korpusy referencyjne. Zgodnie z zasadami ekwiwalencji dynamicznej, powstający w trakcie tłumaczenia tekst przekładu powinien się cechować naturalnością właściwą innym tekstom tworzonym pierwotnie w języku przekładu. Nie zawsze udaje się taką poprawność osiągnąć, głównie ze względu na interferencje z językiem źródłowym, poziom trudności tekstu źródłowego oraz inne czynniki sprzyjające obniżeniu kompetencji językowej tłumacza. W takich wypadkach korpus referencyjny pomaga zweryfikować wybory tłumacza w warstwie leksykalno-gramatycznej oraz stylistycznej przekładu. Zilustrujmy to fragmentem studenckiego tłumaczenia angielskiego tekstu prasowego, donoszącego o wystrzeleniu satelity SMOS[8].

(18.3) O: The SMOS *spacecraft* launched on Monday *to study the Earth's water cycle* has passed a key mission milestone.

(18.4) T: Wystrzelony w poniedziałek *statek kosmiczny SMOS* (ang. Soil Moisture and Ocean Salinity) *mający na celu* obserwację obiegu wody na Ziemi osiągnął już najważniejszy etap swej misji.

Pomijając problem z interpretacją określenia *a key mission milestone*, które oznacza raczej *ważny*, lub *kluczowy* (a nie *najważniejszy*) etap misji, warto zwrócić uwagę na wytłuszczone w powyższym przykładzie fragmenty, które opisują główne zadanie satelity SMOS. W polskiej wersji pojawia się określenie *mający na celu*, które ma w zamierzeniu być dość dynamicznym (niedosłownym) ekwiwalentem angielskiego zdania okolicznikowego celu *to study the Earth's water cycle*. Wydaje się, że sformułowanie *statek kosmiczny ma coś na celu* brzmi niezbyt składnie. Można w takim wypadku po prostu zdecydować się na lepszy odpowiednik, ale w kontekście dydaktyki i analizy tłumaczenia warto się zastanowić, na czym dokładnie polega niezręczność użytego tutaj tłumaczenia. Tabela 18.5 ukazuje przykłady rzeczowników wygenerowanych z NKJP za pomocą narzędzia Kolokator najczęściej występujących przed wyrażeniem *mieć na celu*[9]. Co ciekawe, zdecydowaną większość tych rzeczowników czy też fraz rzeczownikowych można uznać za bliższe lub dalsze hiponimy pojęcia *działanie*, *czynność*, lub *zespół działań* (np. *projekt*). Można tu mówić o pewnym wzorcu leksykalno-gramatycznym, czy też walencyjnym utrwalonego frazeologicznie wyrażenia *mieć na celu*. Typowym jego argumentem jest kategoria semantyczna hiponimu pojęcia *czynność*, do której należy większość grup rzeczownikowych występujących przed tym wyrażeniem. Wzorzec ten ma charakter tendencyjny, a nie kategoryczny. Nietrudno

[8] Tekst dostępny w serwisie BBC, pod adresem: `http://news.bbc.co.uk/2/hi/science/nature/8341400.stm`.

[9] Zob. `http://nkjp.uni.lodz.pl/?q=6jhs7up`.

Tabela 18.5. *mieć na celu* w NKJP

Nr	Kolokacja	Przykład	Źródło
1.	poprawka	Poprawka dwudziesta piąta *ma na celu* poprawienie błędu językowego.	Sprawozdanie z posiedzenia Senatu
2.	zmiana	Wprowadzane przez Microsoft poprzednie zmiany *miały na celu* zwiększenie zasięgu wyszukiwarki i udziału w rynku.	„Dziennik Internautów"
3.	akcja	Akcja *ma na celu* wytępienie komarów przenoszących śmiercionośnego wirusa Zachodniego Nilu.	„Dziennik Polski"
4.	spotkanie	Spotkanie *miało na celu* opracowanie zawartości merytorycznej i organizacji projektu.	„Czas Ostrzeszowski"
5.	sparingi	Sparingi te *miały na celu* sprawdzenie, w jakiej dyspozycji jest każdy z zawodników.	„Nakielski Czas"
6.	program	Program *ma na celu* udrożnienie w stu procentach zbiorników wodnych na Mazowszu.	„Tygodnik Ciechanowski"
7.	zmiany	Wszystkie te zmiany *mają na celu* uzyskanie jak najszybciej najbardziej wydajnego transferu danych.	„Enter"
8.	impreza	Impreza *ma na celu* prezentację kultury alternatywnej i animację przestrzeni miejskiej.	„Polska The Times"
9.	projekt	Projekt *ma na celu* wspieranie rozwoju społeczeństwa informacyjnego oraz gospodarki cyfrowej.	„Nowiny Raciborskie"
10.	wizyta	Moja dzisiejsza wizyta *miała na celu* poznanie i pomoc w rozwiązaniu najważniejszych problemów, z jakimi borykają się w Radomsku policja, straż pożarna i władze samorządowe.	„Gazeta Radomszczańska"

znaleźć wyjątki od powyższej reguły, ale jednocześnie trudno tę tendencję uznać za przypadkową, czy też nieistotną w opisie użycia tego wyrażenia. Na podstawie materiału korpusowego z NKJP można stwierdzić, że w odróżnieniu od rzeczowników oznaczających rodzaj czynności lub działania, *statek kosmiczny* nie jest typowym argumentem walencyjnym wyrażenia *mieć na celu* i chyba głównie

dlatego razi użycie tego sformułowania w tekście tłumaczenia. Z doświadczenia dydaktycznego autora wynika, że tego typu problemy stają się często kością niezgody w dyskusji nad oceną przekładu początkujących tłumaczy. Trudno jest bowiem przekonać studenta translatoryki do przyznania się do subtelnych błędów leksykalno-gramatycznych, zwłaszcza jeżeli popełnione zostały w języku ojczystym tłumacza. Konfrontacja z przykładami użycia danej struktury w korpusie referencyjnym ma w tym względzie dużą wartość dydaktyczną i umacnia świadomość frazeologiczną oraz leksykalno-gramatyczną studentów translatoryki.

18.4. Weryfikacja rejestru ekwiwalentu

Na koniec warto wspomnieć o możliwości użycia korpusu referencyjnego do weryfikacji typowego rejestru rozważanych przez tłumacza ekwiwalentów leksykalnych. Korpusy referencyjne języka oryginału i przekładu są szczególnie przydatne w sytuacji, gdy para potencjalnych ekwiwalentów leksykalnych, lub terminologicznych, mimo bardzo podobnych definicji słownikowych występuje w znacząco różnych typach funkcjonalnych tekstów. Przykładem takiej niepełnej ekwiwalencji dystrybucyjnej może być para wyrazów *pathology* (ang.) oraz *patologia* (pl). Większość przykładów użycia wyrazu *pathology* występujących w korpusie BNC pochodzi z tekstów medycznych lub socjologicznych. Bardzo rzadko słowo to jest używane w potocznej angielszczyźnie mówionej, która w BNC jest reprezentowana[10]. W zbiorze danych NKJP, poza medycznymi i socjologicznymi użyciami tego słowa, można z kolei znaleźć sporo przykładów pochodzących z kontekstów niefachowych i nieformalnych, na przykład z języka mówionego oraz z list i forów internetowych, na których toczy się (dość często zwulgaryzowana) dyskusja na temat problemów społecznych, na przykład[11,12]:

(18.5) Jak patrzę na punka to nie wiem skąd się tacy ludzie biorą. *Patologia*. (`www.forumowisko.pl`)

(18.6) Może w końcu ludzie tacy jak mój brat, którzy są w drugiej klasie liceum wiedzieliby przynajmniej ile to jest 25% z szesnastu... Przecież teraz to jest normalnie *patologia*. (`www.forumowisko.pl`)

(18.7) Kwalifikacje nie mają znaczenia. Liczą się koneksje. To pierwsza *patologia*. (Usenet – `pl.soc.polityka`)

[10] Zob. `http://nkjp.uni.lodz.pl/?q=4ccnaym`.

[11] Zob. `http://nkjp.uni.lodz.pl/?q=6bcwsqy`.

[12] Trzeba tu zaznaczyć, że w BNC fora dyskusyjne i inne teksty internetowe w ogóle nie są reprezentowane, częściowo ze względu na czas powstania tego korpusu. Niemniej jednak hipotezę o niesymetryczności rejestrów występowania polskiej *patologii* i angielskiej *pathology* potwierdzają bardziej aktualne korpusy referencyjne angielszczyzny, np. Korpus Współczesnej Angielszczyzny Amerykańskiej (`www.americancorpus.org`).

Nie jest to oczywisty przypadek „fałszywych przyjaciół" tłumacza, ponieważ podstawowe znaczenia *patologii* oraz *pathology* są podobne, a niekiedy wprost identyczne (*nieprawidłowe zjawiska występujące w życiu społecznym*[13]), a rejestry ich użycia do pewnego stopnia się pokrywają (użycie tego wyrazu w angielszczyźnie wydaje się być dużo węższe). Widoczne jednak w korpusach referencyjnych asymetrie w dystrybucji obu wyrazów nakazują zachować szczególną ostrożność, zwłaszcza podczas tłumaczenia na język angielski niefachowych użyć wyrazu *patologia*.

Chociaż NKJP nie jest w zamierzeniu korpusem zrównoważonym diachronicznie, to można go w wielu wypadkach użyć do weryfikacji historycznej danego wyrazu czy frazy, co pozwala uniknąć użycia anachronicznego ekwiwalentu w pewnych kontekstach. Na przykład jedne z najwcześniejszych w NKJP użyć czasownika *ściemniać* w znaczeniu *kłamać*, *nie mówić prawdy*, *udawać niewiedzę* pochodzą z 1996 roku[14]:

(18.8) - Dyrektor „*ściemniał*" - Darek slangiem określa zachowanie pedagoga. - Udawał, że nie słyszy i kazał mi odebrać wyprawkę: koc, kołdrę, poduszkę, komplet talerzy i sztućce. („Gazeta Wyborcza" z 27.12.1996)

(18.9) Mówiłem, może byśmy pojechali do mnie. Żadnego *ściemniania*, wszystko jasne. (*Chłopaki nie płaczą*, 1996)

Nie są to oczywiście ostateczne dowody na pierwsze użycie tego znaczenia, ale warto posiadać takie informacje, tłumacząc na język polski tekst, który powstał na długo przed rokiem 1996, zwłaszcza jeżeli tłumacz decyduje się na równoważny diachronicznie przekład. Innymi słowy, tłumacz może sprawdzić, czy proponowany przez niego ekwiwalent nie jest anachroniczny w stosunku do tekstu oryginału.

18.5. Podsumowanie

Narodowy Korpus Języka Polskiego to duży korpus referencyjny i wraz z dostępnymi publicznie narzędziami do jego przeszukiwania jest cennym elementem w warsztacie tłumacza. Może on niewątpliwie stanowić ważne uzupełnienie słowników przy weryfikacji frazeologii przekładu. Dzięki temu tłumaczenie może być nie tylko wierniejsze oryginałowi, ale także poprawniejsze gramatycznie i przystępniejsze w odbiorze. Ważnym zastosowaniem NKJP w pracy tłumacza jest także weryfikacja poprawności leksykalno-gramatycznej przekładu, zilustrowana

[13] Zob. `http://sjp.pwn.pl/szukaj/patologia`.

[14] Zob. `http://nkjp.uni.lodz.pl/?q=6hyofba`.

w tym rozdziale przykładem użycia relacji walencyjnych czasowników i zleksykalizowanych fraz czasownikowych. Dzięki różnorodności rejestrów, typów funkcjonalnych tekstów, oraz, do pewnego stopnia, okresów historycznych reprezentowanych w NKJP, korpus ten stanowi także ważne źródło przydatnych w przekładzie informacji o cechach dystrybucyjnych wyrazów, fraz oraz konstrukcji gramatycznych w polszczyźnie. Zasoby NKJP są już teraz wykorzystywane nie tylko w dydaktyce translacji, ale też przez profesjonalnych tłumaczy.

Bibliografia

Acedański S. (2010). *A morphosyntactic Brill tagger for inflectional languages*. W: Loftsson H., Rögnvaldsson E. i Helgadóttir S. (red.), *Advances in Natural Language Processing: Proceedings of the 7th International Conference on Natural Language Processing, IceTAL 2010, Reykjavík, Iceland*, t. 6233 serii *Lecture Notes in Artificial Intelligence*, s. 3–14, Heidelberg. Springer-Verlag.

Acedański S. i Gołuchowski K. (2009). *A morphosyntactic rule-based Brill tagger for Polish*. W: Kłopotek M. A., Przepiórkowski A., Wierzchoń S. T. i Trojanowski K. (red.), *Recent Advances in Intelligent Information Systems*, s. 67–76. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Acedański S. i Przepiórkowski A. (2010). *Towards the adequate evaluation of morphosyntactic taggers*. W: *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010); Poster Session*, s. 1–8, Beijing.

Agirre E. i Edmonds P., red. (2006). *Word Sense Disambiguation: Algorithms and Applications*, t. 33 serii *Text, Speech and Language Technology*. Springer, Dordrecht.

Alex B., Haddow B. i Grover C. (2007). *Recognising nested named entities in biomedical text*. W: *Proceedings of the BioNLP Workshop at ACL 2007*, s. 65–72, Praga.

Atkins S., Clear J. H. i Ostler N. (1992). *Corpus design criteria*. „Journal of Literary and Linguistic Computing", 7(1), s. 1–16.

Baker M. (1992). *In Other Words: a Coursebook on Translation*. Routledge, Londyn.

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

Bańko M. (2001). *Z pogranicza leksykografii i językoznawstwa*. Uniwersytet Warszawski, Wydział Polonistyki, Warszawa.

Bańko M., red. (2005). *Multimedialny słownik szkolny PWN*. Wydawnictwo Naukowe PWN. Wersja 1.0, płyta CD.

Bańko M. (2010). *Jedność w wielości. Cechy konstytutywne i typologiczne słowników*. „Poradnik Językowy", 4, s. 5–25.

Bański P. i Przepiórkowski A. (2009). *Stand-off TEI annotation: the case of the National Corpus of Polish*. W: *Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, s. 64–67, Singapur.

Bański P. i Przepiórkowski A. (2010). *The TEI and the NCP: the model and its application*. W: *LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, Valletta, Malta. ELRA.

Bartmiński J., red. (2001). *Współczesny język polski*. Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej, Lublin.

Becker M., Drożdżyński W., Krieger H.-U., Piskorski J., Schäfer U. i Xu F. (2002). *SProUT – Shallow Processing with Typed Feature Structures and Unification*. W: *Proceedings of ICON 2002, Mumbay, India*.

Bejček E. i Straňák P. (2010). *Annotation of multiword expressions in the Prague Dependency Treebank*. „Language Resources and Evaluation", 44(1–2), s. 7–21.

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

Biber D. (1993). *Representativeness in corpus design*. „Literary and Linguistic Computing", 8(4), s. 243–257.

Biber D. (2004). *Lexical bundles in academic speech and writing*. W: Lewandowska-Tomaszczyk B. (red.), *The proceedings of Practical Applications in Language and Computers PALC 2003*, Frankfurt nad Menem. Peter Lang.

Bień J. S. (1991). *Koncepcja słownikowej informacji morfologicznej i jej komputerowej weryfikacji*, t. 383 serii *Rozprawy Uniwersytetu Warszawskiego*. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa. `http://bc.klf.uw.edu.pl/12/`.

Bień J. S. i Saloni Z. (1982). *Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna)*. „Prace Filologiczne", XXXI, s. 31–45.

Böhmová A., Hajič J., Hajičová E. i Hladká B. (2003). *The Prague Dependency Treebank: Three-level annotation scenario*. W: Abeillé A. (red.), *Treebanks: Building and Using Parsed Corpora*, t. 20 serii *Text, Speech and Language Technology*, s. 103–127. Kluwer, Dordrecht.

Bolinger D. (1965). *The atomization of meaning*. „Language", 41(4), s. 555–573.

Brill E. (1992). *A simple rule-based part of speech tagger*. W: *Proceedings of the 3rd Applied Natural Language Processing Conference*, s. 152–155, Trento. ACL.

Brill E. (1994). *Some advances in transformation-based part of speech tagging*. W: *Proceedings of AAAI-94*, s. 722–727.

Buczyński A. i Przepiórkowski A. (2009). *Spejd: A shallow processing and morphological disambiguation tool*. W: Vetulani Z. i Uszkoreit H. (red.), *Human Language Technology: Challenges of the Information Society*, t. 5603 serii *Lecture Notes in Artificial Intelligence*, s. 131–141. Springer-Verlag, Berlin.

Budiscak I., Piskorski J. i Ristov S. (2009). *Compressing gazetteers revisited*. W: *Proceedings of the 8th International Workshop on Finite-State Methods and Natural Language Processing, Pretoria, South Africa*.

Burnard L. i Bauman S., red. (2008). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford. `http://www.tei-c.org/Guidelines/P5/`.

Čermák F. (2001). *Language corpora: The Czech case*. W: Matoušek V., Mautner P., Moucek R. i Tauser K. (red.), *Text, Speech and Dialogue: 4th International Conference, TSD 2001, Zelezna Ruda, Czech Republic, September 2001*, t. 2166 serii *Lecture Notes in Artificial Intelligence*, s. 21–30, Berlin. Springer-Verlag.

Chinchor N. (1997). *MUC-7 Named Entity Task Definition*. W: *Proc. of MUC-7*.

Cohen J. (1960). *A coefficient of agreement for nominal scales*. „Educational and Psychological Measurement", 20, s. 37–46.

Czerepowicka M. (2005). *Opis lingwistyczny wyrażeń niestandardowych składniowo typu* na lewo, do dziś, po trochu, na zawsze *we współczesnym języku polskim*. Rozprawa doktorska, Uniwersytet Warmińsko-Mazurski, Olsztyn.

Daciuk J. i Piskorski J. (2006). *Gazetteer compression technique based on substructure recognition*. „Advances in Soft Computing", 35.

Derwojedowa M. i Rudolf M. (2003). *Czy Burkina to dziewczyna i co o tym sądzą ich królewskie mości, czyli o jednostkach leksykalnych pewnego typu*. „Poradnik Językowy", 5, s. 39–49.

Desmet B. i Hoste V. (2010). *Towards a balanced Named Entity corpus for Dutch*. W: *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Drożdżyński W., Krieger H.-U., Piskorski J., Schäfer U. i Xu F. (2004). *Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications*. „Künstliche Intelligenz", 1/04.

Dubisz S., red. (2003). *Uniwersalny słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Dunaj B. (1987). *Czas przyszły czasowników niedokonanych w polszczyźnie – uzus i norma*. „Język Polski", LXVII, s. 9–19.

Dziwirek K. i Lewandowska-Tomaszczyk B. (2010). *Complex Emotions and Grammatical Mismatches: A Contrastive Corpus-Based Study*. Mouton de Gruyter, Berlin.

Finkel J. R. i Manning C. D. (2009). *Joint parsing and named entity recognition*. W: *Proceedings of HLT-NAACL 2009*, s. 326–334. ACL.

Gajda S., red. (1995). *Przewodnik po stylistyce polskiej*. Uniwersytet Opolski. Instytut Filologii Polskiej, Opole.

Givón T. (1983). *Topic continuity in discourse: a quantitative cross-language study.* John Benjamins, Amsterdam.

Głowiński M., Okopień-Sławińska A. i Sławiński J. (1986). *Zarys teorii literatury*. Wydawnictwa Szkolne i Pedagogiczne, Warszawa.

Goźdź-Roszkowski S., red. (2011). *Explorations across Languages and Corpora: PALC 2009*, Frankfurt nad Menem. Peter Lang.

Górski R. L. (2008). *Diateza nacechowana w polszczyźnie. Studium korpusowe*. Lexis, Kraków.

Górski R. L. i Łaziński M. (2010). *Wzór stylu i wzór na styl. Zróżnicowanie stylistyczne tekstów Narodowego Korpusu Języka Polskiego*. W: Milewska-Stawiany M. i Rogowska-Cybulska E. (red.), *Polskie języki. O językach zawodowych i środowiskowych. Materiały VII Forum Kultury Słowa*, s. 41–55, Gdańsk. Wydawnictwo Uniwersytetu Gdańskiego.

Grzegorczyk P. (1967). *Index lexicorum Poloniae. Bibliografia słowników polskich*. Wydawnictwo Naukowe PWN, Warszawa.

Hajič J. (1997). *Probabilistic and rule-based tagger of an inflective language — a comparison*. W: *Proceedings of the 5th Applied Natural Language Processing Conference*, s. 111–118, Washington, D.C. ACL.

Hajnicz E., Murzynowski G. i Woliński M. (2008). *ANOTATORNIA – lingwistyczna baza danych*. W: *Materiały V konferencji naukowej InfoBazy 2008, Systemy * Aplikacje * Usługi*, s. 168–173, Gdańsk. Centrum Informatyczne TASK, Politechnika Gdańska.

Halliday M. (2004). *Lexicology and Corpus Linguistics. An Introduction.* Continuum, Londyn.

Hebal-Jezierska M. (2008). *Wariantywność końcówek fleksyjnych rzeczowników męskich żywotnych w języku czeskim*. Uniwersytet Warszawski, Wydział Polonistyki, Warszawa.

Hinrichs E., Kübler S., Naumann K. i Zinsmeister H. (2005). *Recent developments in linguistic annotations of the TüBa-D/Z treebank*. W: *27th Annual Meeting of the German Linguistic Association*, Kolonia, Niemcy.

Holmes-Higgin P., Ahmad K. i Abidi S. S. R. (1994). *A description of texts in a corpus: 'virtual' and 'real' corpora*. W: Martin W., Meijs W. i Moerland M. (red.), *EURALEX 1994 Proceedings. Papers submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, The Netherlands*, s. 390–402, Amsterdam. Vrije Universiteit.

Ide N., Bonhomme P. i Romary L. (2000). *XCES: An XML-based standard for linguistic corpora*. W: *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, s. 825–830, Ateny. ELRA.

Ide N. i Véronis J. (1993). *Background and context for the development of a Corpus Encoding Standard*. EAGLES Working Paper, `http://www.cs.vassar.edu/CES/CES3.ps.gz`.

ISO:24610-1 (2005). *Language resource management – feature structures – part 1: Feature structure representation*. ISO/DIS 24610-1, 2005-10-20.

Janus D. i Przepiórkowski A. (2007). *Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora*. W: Waliński J., Kredens K. i Goźdź-Roszkowski S. (red.), *The proceedings of Practical Applications in Language and Computers PALC 2005*, Frankfurt nad Menem. Peter Lang.

Karłowicz J. (1876). *Przyczynki do projektu wielkiego słownika polskiego*. „Rozprawy i sprawozdania z posiedzeń Wydziału Filologicznego Akademii Umiejętności", 4, s. XIV–XCIV. Kraków.

Karwańska D. i Przepiórkowski A. (2011). *On the evaluation of two Polish taggers*. W: Goźdź-Roszkowski (2011), s. 105–113.

Kilgarriff A., Rychly P., Smrž P. i Tugwell D. (2008). *The Sketch Engine*. W: Fontenelle T. (red.), *Practical Lexicography. A Reader*, s. 297–306. Oxford University Press, Oxford. Pierwodruk w: G. Williams, S. Vessier, red., *Euralex 2004 Proceedings*. Univeristé de Bretagne-Sud, Lorient, s. 105–116.

Klemensiewicz Z. (1982). *Składnia, stylistyka, pedagogika językowa*. Wydawnictwo Naukowe PWN, Warszawa.

Klemensiewicz Z. (1985). *Historia języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Kosyl C. (2001). *Nazwy osobowe*. W: Bartmiński (2001), s. 431–445.

Kubiak-Sokół A. i Łaziński M., red. (2007). *Słownik nazw miejscowości i mieszkańców*. Wydawnictwo Naukowe PWN, Warszawa.

Kučera K. (2002). *The Czech National Corpus: principles, design, and results*. „Literary and Linguistic Computing", 17(2), s. 245–257.

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

Kupść A. (1999). *Haplology of the Polish reflexive marker*. W: Borsley R. D. i Przepiórkowski A. (red.), *Slavic in Head-Driven Phrase Structure Grammar*, s. 91–124. CSLI Publications, Stanford, CA.

Kurcz I., Lewicki A., Sambor J. i Woronczak J. (1974a). *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom I. Teksty popularnonaukowe*. Maszynopis, Uniwersytet Warszawski.

Kurcz I., Lewicki A., Sambor J. i Woronczak J. (1974b). *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom II. Drobne wiadomości prasowe*. Maszynopis, Uniwersytet Warszawski.

Kurcz I., Lewicki A., Sambor J. i Woronczak J. (1976). *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom IV. Proza artystyczna*. Maszynopis, Uniwersytet Warszawski.

Kurcz I., Lewicki A., Sambor J. i Woronczak J. (1977). *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom V. Dramat artystyczny*. Maszynopis, Uniwersytet Warszawski.

Kurcz I., Lewicki A., Sambor J., Szafran K. i Woronczak J. (1990). *Słownik frekwencyjny polszczyzny współczesnej*. Wydawnictwo Instytutu Języka Polskiego PAN, Kraków.

Kurcz I. i Polkowska A., red. (1990). *Interakcyjne i autonomiczne przetwarzanie informacji językowych. Na przykładzie procesu rozumienia tekstu czytanego na głos*. Zakład narodowy im. Ossolińskich, Wrocław.

Lakoff G. i Johnson M. (1980). *Metaphors We Live By*. Chicago University Press, Chicago, IL.

Lakoff G. i Johnson M. (2010). *Metafory w naszym życiu*. Ossolineum, Wrocław.

Lewandowska-Tomaszczyk B. (2004). *Podstawy językoznwstwa korpusowego*. Wydawnicwto Uniwersytetu Łódzkiego, Łódź.

Lewandowska-Tomaszczyk B. (2012). *Approximative spaces and tolerance threshold in communication*. W: Lewandowska-Tomaszczyk B. i Thelen M. (red.), *Translation and Meaning Vol. 10*. Zuyd University, Maastricht School of International Communication, Maastricht.

Lewicki A., Masłowski W., Sambor J. i Woronczak J. (1975). *Słownictwo współczesnego języka polskiego. Listy frekwencyjne. Tom III. Publicystyka*. Maszynopis, Uniwersytet Warszawski.

Łaziński M. i Szewczyk M. (2006). *Słowa klucze w semantyce i statystyce. Słowa tygodnia „Rzeczpospolitej"*. „Biuletyn Polskiego Towarzystwa Językoznawczego", LXII, s. 57–68.

Makarenko V. (2001). *Nie tylko liczyć, trzeba też rozumieć*. Gazeta Wyborcza, wydanie 288, s. 22.

Marcińczuk M. i Piasecki M. (2011). *Statistical proper name recognition in Polish economic texts*. „Control and Cybernetics", 40(2), s. 393–418.

Matuszczyk B. (2006). *Słownik języka polskiego S.B. Lindego. Warsztat leksykografa*. Katolicki Uniwersytet Lubelski, Lublin.

Maurel D. (2008). *Prolexbase. A multilingual relational lexical database of proper names*. W: *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marakesz. ELRA.

Megyesi B. (1999). *Improving Brill's POS tagger for an agglutinative language*. W: *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, s. 275–284, College Park, MD. ACL.

Michalski J. (1961). *Dzieje wydania* Słownika *Lindego*. „Studia i materiały z dziejów nauki polskiej. Seria A: historia nauk społecznych", 4, s. 5–39.

Milewska B. (2003a). *Przyimki wtórne we współczesnej polszczyźnie*. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Milewska B. (2003b). *Słownik polskich przyimków wtórnych*. Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Miłkowski M. i Lipski J. (2011). *Using SRX standard for sentence segmentation*. W: Vetulani Z. (red.), *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6–8, 2009, Revised Selected Papers*, t. 6562 serii *Lecture Notes in Artificial Intelligence*, s. 172–182, Berlin. Springer-Verlag.

Młodzki R. i Przepiórkowski A. (2009). *The WSD development environment*. W: Vetulani (2009), s. 185–189.

Mykowiecka A., Marciniak M. i Rabiega-Wiśniewska J. (2008). *Proper names in Polish dialogs*. W: *Proceedings of the IIS'2008 Workshop on Spoken Language Understanding and Dialogue Systems*, Zakopane. Springer-Verlag.

Ngai G. i Florian R. (2001). *Transformation-based learning in the fast lane*. W: *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, s. 1–8.

Nida E. A. i Taber C. (1969). *The Theory and Practice of Translation*. Brill, Leiden.

Olkiewicz J. (1988). *Od A do Z, czyli o encyklopediach i encyklopedystach*. Ludowa Spółdzielnia Wydawnicza, Warszawa.

Pajas P. i Štěpánek J. (2008). *Recent advances in a feature-rich framework for treebank annotation*. W: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, s. 673–680, Manchester.

Pęzik P. (2011). *Providing corpus feedback for translators with the PELCRA search engine for NKJP*. W: Goźdź-Roszkowski (2011), s. 135–144.

Piasecki M. (2006). *Hand-written and automatically extracted rules for Polish tagger*. W: Sojka P., Kopeček I. i Pala K. (red.), *Text, Speech and Dialogue: 9th International Conference, TSD 2006, Brno, Czech Republic, September 2006*, t. 4188 serii *Lecture Notes in Artificial Intelligence*, s. 205–212, Berlin. Springer-Verlag.

Piasecki M. (2007). *Polish tagger TaKIPI: Rule based construction and optimisation*. „Task Quarterly", 11(1–2), s. 151–167.

Piper D. (2003). *Współczesna polska ortografia w szkole na przykładzie gimnazjum*. Wydawnictwo Uniwersytetu Warmińsko-Mazurskiego, Olsztyn.

Piskorski J. (2005). *Named-Entity Recognition for Polish with SProUT*. W: *LNCS Vol 3490: Proceedings of IMTCI 2004, Warsaw, Poland*.

Piskorski J., Homola P., Marciniak M., Mykowiecka A., Przepiórkowski A. i Woliński M. (2004). *Information Extraction for Polish Using the SProUT Platform*. W: *Proceedings of IIS'04, Zakopane, Poland*.

Piskorski J., Sydow M. i Kupść A. (2007). *Lemmatization of Polish person names*. W: Piskorski J., Pouliquen B., Steinberger R. i Tanev H. (red.), *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing at ACL 2007*, s. 27–34, Praga.

Polański K., red. (1993). *Encyklopedia językoznawstwa ogólnego*. Ossolineum, Wrocław.

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

Przepiórkowski A. (2006). *O inherentnej liczbie mnogiej liczebników* ćwierć, pół *i* półtora. „Poradnik Językowy", 9, s. 78–87.

Przepiórkowski A. (2008). *Powierzchniowe przetwarzanie języka polskiego*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Przepiórkowski A. (2009a). *A comparison of two morphosyntactic tagsets of Polish*. W: Koseska-Toszewa V., Dimitrova L. i Roszko R. (red.), *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, s. 138–144, Warszawa.

Przepiórkowski A. (2009b). *TEI P5 as an XML standard for treebank encoding*. W: Passarotti M., Przepiórkowski A., Raynaud S. i Van Eynde F. (red.), *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, s. 149–160, Milan, Italy.

Przepiórkowski A. i Bański P. (2009). *Which XML standards for multilevel corpus annotation?* W: Vetulani (2009), s. 245–250.

Przepiórkowski A. i Bański P. (2011). *XML text interchange format in the National Corpus of Polish*. W: Goźdź-Roszkowski (2011), s. 55–65.

Przepiórkowski A., Kupść A., Marciniak M. i Mykowiecka A. (2002). *Formalny opis języka polskiego: Teoria i implementacja*. Akademicka Oficyna Wydawnicza EXIT, Warszawa.

Przepiórkowski A. i Murzynowski G. (2011). *Manual annotation of the National Corpus of Polish with Anotatornia*. W: Goźdź-Roszkowski (2011), s. 95–103.

Przepiórkowski A. i Woliński M. (2003). *A flexemic tagset for Polish*. W: *Proceedings of Morphological Processing of Slavic Languages, EACL 2003*, s. 33–40, Budapeszt.

Ptaszyk M. (2007). *„Słownik języka polskiego" Samuela Bogumiła Lindego*. Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

Quasthoff U., Richter M. i Wolff C. (2002). *Wörter des Tages – Tagesaktuelle wissensbasierte Analyse und Visualisierung von Zeitungen und Newsdiensten*. W: Hammwöhner R., Wolff C. i Womser-Hacker C. (red.), *Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information. Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002), Regensburg, 8.-10. Oktober 2002*, s. 369–372, Konstanz. UVK Verlagsgesellschaft.

Ramshaw L. A. i Marcus M. P. (1995). *Text chunking using transformation-based learning*. W: *Proceedings of the Third Workshop on Very Large Corpora*, s. 82–94, Cambridge, MA. ACL.

Rospond S. (1992–1994). *Słownik nazwisk śląskich*, t. I–II. Ossolineum, Wrocław.

Rymut K., red. (1980). *Nazwy miast polskich*. Ossolineum, Wrocław.

Rymut K., red. (1992–1994). *Słownik nazwisk współcześnie w Polsce używanych*, t. I–X. Wydawnictwo Instytutu Języka Polskiego PAN, Kraków.

Rymut K., red. (1995). *Słownik imion współcześnie w Polsce używanych*, t. I-X. Wydawnictwo Instytutu Języka Polskiego PAN, Kraków.

Rymut K. (2002). *Dictionary of Surnames in Current Use in Poland at the Beginning of the 21st Century*. Wydawnictwo Instytutu Języka Polskiego PANi Polish Genealogical Society of America, Kraków, Chicago.

Rymut K., red. (2008). *Nazwy wodne Polski*. Wydawnictwo Instytutu Języka Polskiego PAN, Kraków.

Rzetelska-Feleszko E. (2001a). *Nazwy geograficzne*. W: Bartmiński (2001), s. 411–429.

Rzetelska-Feleszko E. (2001b). *Nazwy własne*. W: Bartmiński (2001), s. 405–410.

Rzetelska-Feleszko E., red. (2005). *Polskie nazwy własne*. Wydawnictwo Instytutu Języka Polskiego PAN, Kraków.

Saeed J. (2009). *Semantics*. Wiley-Blackwell, Malden, MA, wyd. III.

Saloni Z. (1988). *O tzw. formach nieosobowych [rzeczowników] męskoosobowych we współczesnej polszczyźnie*. „Biuletyn Polskiego Towarzystwa Językoznawczego", XLI, s. 155–166.

Saloni Z. (2001). *Czasownik polski. Odmiana, słownik*. Wiedza Powszechna, Warszawa.

Saloni Z., Gruszczyński W., Woliński M. i Wołosz R. (2007). *Słownik gramatyczny języka polskiego*. Wiedza Powszechna, Warszawa.

Saloni Z. i Świdziński M. (2001). *Składnia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa, wyd. V.

Sang E. F. T. K. i de Meulder F. (2003). *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. W: *Proceedings of CoNLL-2003, Edmonton, Canada*, s. 142–147.

Savary A. i Piskorski J. (2011). *Language resources for Named Entity annotation in the National Corpus of Polish*. „Control and Cybernetics", 40(2), s. 361–391.

Savary A., Rabiega-Wiśniewska J. i Woliński M. (2009). *Inflection of Polish multi-word proper names with Morfeusz and Multiflex*. W: Marciniak M. i Mykowiecka A. (red.), *Aspects of Natural Language Processing*, t. 5070 serii *Lecture Notes in Computer Science*, s. 111–141. Springer-Verlag, Berlin.

Scott M. (2007). *PC analysis of key words and key key words*. W: Teubert W. i Krishnamurthy R. (red.), *Corpus Linguistics: Critical Concepts in Linguistics*, s. 303–337. Routledge, Nowy Jork.

Sekine S., Sudo K. i Nobata C. (2002). *Extended Named Entity hierarchy*. W: *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*, Las Palmas. ELRA.

Sinclair J., red. (1987). *Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. Collins, Londyn.

Slavicorp (2012). *Slavic corpus linguistics*. „Prace Filologiczne", **LXIV** (w druku).

Straus G. i Wolff K. (1996a). *Czytanie i kupowanie książek w Polsce w 1994 r.: (raport z badań)*. Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Straus G. i Wolff K. (1996b). *Polacy i książki: społeczna sytuacja książki w Polsce 1992*. Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Straus G. i Wolff K. (1998). *Zainteresowanie książką w społeczeństwie polskim w 1996 r.: (raport z badań)*. Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Straus G. i Wolff K. (2000). *Czytać, nie czytać…, kupować, nie kupować…: sytuacja książki w społeczeństwie polskim w 1998 r.* Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Straus G. i Wolff K. (2002). *Sienkiewicz, Mickiewicz, Biblia, harlequiny…: społeczny zasięg książki w Polsce w 2000 roku*. Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Straus G. i Wolff K. (2004). *Kiążka na początku wieku: społeczny zasięg książki w Polsce w 2002 roku*. Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Straus G. i Wolff K. (2006). *Czytanie, kupowanie, wypożyczanie: społeczny zasięg książki w Polsce w 2004 roku*. Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Straus G., Wolff K. i Wierny S. (2008). *Czytanie, kupowanie, surfowanie: społeczny zasięg książki w Polsce w 2006 roku*. Biblioteka Narodowa. Instytut Książki i Czytelnictwa, Warszawa.

Stubbs M. (1996). *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Blackwell Publishers, Oxford.

Sutton C. i McCallum A. (2007). *An introduction to conditional random fields for relational learning*. W: Getoor L. i Taskar B. (red.), *Introduction to Statistical Relational Learning*, chapter 4. The MIT Press.

Szałkiewicz Ł. (2010). *Chamy posły i zuchy doktory — głos w sprawie deprecjatywności*. „LingVaria", 9, s. 219–232.

Szkiłądź H. (1997). *Redakcja Słowników Języka Polskiego*. W: *Alfabet PWN. Ludzie, książki, lata, wspomnienia*, s. 326–336. Wydawnictwo Naukowe PWN, Warszawa.

Szul R. (2009). *Język Naród Państwo. Język jako zjawisko polityczne*. Wydawnictwo Naukowe PWN.

Szymczak M., red. (1994). *Słownik języka polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Świdziński M. (1996). *Własności składniowe wypowiedników polskich*. Dom Wydawniczy Elipsa, Warszawa.

Teubert W. (2005). *My version of corpus linguistics*. „International Journal of Corpus Linguistics", 10, s. 1–13.

The Economist (2009). *A heated debate. Why political orthodoxy must not silence scientific argument*.

Twardzik W., red. (2006). *Biblioteka zabytków polskiego piśmiennictwa średniowiecznego*. Wydawnictwo Instytutu Języka Polskiego PAN, Kraków.

Urbańczyk S., red. (1992). *Encyklopedia języka polskiego*. Ossolineum, Wrocław.

Vetulani Z., red. (2009). *Proceedings of the 4th Language & Technology Conference*, Poznań, Poland.

Wallach H. M. (2004). *Conditional random fields: An introduction*. Technical Report MS-CIS-04-21, University of Pennsylvania.

Waszczuk J., Głowińska K., Savary A., Przepiórkowski A. i Lenart M. (2011). *Annotation Tools for Syntax and Named Entities in the National Corpus of Polish*. „International Journal of Data Mining, Modelling and Management". W druku.

Wilcock G. (2009). *Introduction to Linguistic Annotation and Text Analytics*. Morgan & Claypool.

Witten I. H. i Frank E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, wyd. II. `http://www.cs.waikato.ac.nz/ml/weka/`.

Wojtak M. (2008). *Genologia tekstów użytkowych*. W: Ostaszewska D. i Cudak R. (red.), *Polska genologia lingwistyczna*, s. 339–352, Warszawa. Wydawnictwo Naukowe PWN.

Woliński M. (2003). *System znaczników morfosyntaktycznych w korpusie IPI PAN*. „Polonica", XXII–XXIII, s. 39–55.

Woliński M. (2006). *Morfeusz — a practical tool for the morphological analysis of Polish*. W: Kłopotek M. A., Wierzchoń S. T. i Trojanowski K. (red.), *Intelligent Information Processing and Web Mining*, Advances in Soft Computing, s. 503–512. Springer-Verlag, Berlin.

Woliński M. i Przepiórkowski A. (2001). *Projekt anotacji morfosynktaktycznej korpusu języka polskiego*. Prace IPI PAN 938, Instytut Podstaw Informatyki PAN, Warszawa.

Yarowsky D. (1993). *One sense per collocation*. W: *Proceedings of the workshop on Human Language Technology*, HLT'93, s. 266–271, Stroudsburg, PA. Association for Computational Linguistics.

Załuski J. (1976). *Wspomnienia*. Wydawnictwo Literackie, Kraków. Wstęp i opracowanie: Anna Palarczykowa.

ZSRR (1978). *Konstytucja Związku Socjalistycznych Republik Radzieckich*. Książka i Wiedza.

Żmigrodzki P., Bańko M., Dunaj B. i Przybylska R. (2007). *Koncepcja* Wielkiego słownika języka polskiego — *przybliżenie drugie*. W: Żmigrodzki P. i Przybylska R. (red.), *Nowe studia leksykograficzne*, s. 9–21. Lexis, Kraków.

Żmigrodzki P., Ulitzka E. i Nowak T. (2005). *O świadomości leksykograficznej kandydatów na polonistów (na podstawie badań ankietowych)*. „Poradnik Językowy", 5, s. 3–21.

Wydawnictwo Naukowe PWN SA
Wydanie pierwsze
Skład i łamanie w systemie LaTeX: Adam Przepiórkowski, Warszawa
Druk i oprawa: Wrocławska Drukarnia Naukowa PAN Sp. z o.o.