„Bohemistyka” 2008, nr 1–4, ISSN 1642–9893

Dariusz TKACZEWSKI
Katowice – Ostrawa

Český národní korpus – internetowe źródło standaryzacji i weryfikacji języka czeskiego oraz nowoczesne narzędzie dydaktyczne

Czy tego chcemy, czy nie, stajemy się społeczeństwem informatycznym w coraz większym zakresie. Licząca sobie prawie 30 lat globalna sieć Internetu to nie tylko nieprzebrany informator – (strony www, źródło wiedzy o świecie i ludziach w prawie wszystkich językach), natychmiastowy komunikator (korespondencja za pośrednictwem łączności mailowej i programów typu Gadu-Gadu), to jakby wszechobecny makler, urzędnik bankowy czy sprzedawca sklepowy. Internet powoli staje się tez skutecznym i zawsze podręcznym nauczycielem, konsultantem oraz „strażnikiem” standardu i poprawności językowej wielu języków nowożytnych, w tym współczesnej czeszczyzny. Z biegiem czasu coraz większą popularnością wśród użytkowników języków narodowych dbających o ich poprawność stają się słowniki internetowe, czyli odpowiednie witryny/strony www. przejmują rolę purystycznych słowników języka współczesnego (poprawnego języka, ortograficznego, frekwencyjnego, a tergo itp.). Takie leksykony wirtualne uzupełniają te klasyczne słowniki w formie książki i multimedialnej płyty CD, unowocześniają metody konsultacji językowej oraz dostępu do standardu językowego, poprzez możliwość weryfikacji danej jednostki językowej natychmiast, w czasie rzeczywistym, *on line*.

W odniesieniu do języka współczesnego – zwłaszcza potocznego i mówionego – rolę taką w coraz większej mierze pełnią również tzw. narodowe korpusy językowe, doskonale dokumentujące też płaszczyzny stylistyczne języków ogólnonarodowych. W cyberprzestrzeni Internetu działa obecnie ponad 30 korpusów[1], z tego aż 15 korpusów języków słowiańskich[2] – w tej liczbie dwa polskie (IPI PAN[3], Korpusu

[1] Wśród korpusów języków niesłowiańskich można wymienić: British National Corpus (najstarszy), Collins Cobuild - The Bank of English, Lietuvių kalbos tekstynas, Magyar Nemzeti Szövegtár, AC/DC Portuguese corpora, American National Corpus, WordNet®, SUSANNE Corpus, The Penn Treebank Project, IDS - Korpora der geschriebenen Sprache, Thesaurus Indogermanischer Text- und Sprachmaterialien TITUS, EESTI ÕIGUSAKTID (Korpus estońskich tekstów prawniczych), The Corpus of Spoken Israeli Hebrew, Negr@ corpus (Syntactically Annotated Corpus of German Newspaper Texts), SPRÅKBANKEN (En sprÅklig referensdatabas vid Göteborgs universitet).

[2] Korpusy narodowe języków słowiańskich: Slovenský národný korpus, Korpus DIALOG 0.1 ÚJČ AV ČR, Korpus IPI PAN, Korpus Języka Polskiego Wydawnictwa Naukowego PWN, Korpus slovenskega jezika FIDA, beseda – Besedilni korpus na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU, nova beseda, Korpus bosanskih tekstova na Univerzitetu u Oslu, Corpus of Serbian Language, Bulgarian Treebank, Corpus Cyrillo-Methodianum Helsingiense An Electronic Corpus of Old Church Slavonic Texts, Russian Corpora in Tübingen, Национальный корпус русского языка, Компьютерный корпус текстов русских газет конца XX-ого века, Upper Sorbian Text Corpus – Hornjoserbski tekstowy korpus.

[3] Korpus IPI PAN jest dużym (obecnie ponad 250 000 000 segmentów), anotowanym morfosyntaktycznie, publicznie dostępnym korpusem języka polskiego, stworzonym przez Zespół Inżynierii Lingwistycznej w Instytucie Podstaw Informatyki PAN (IPI PAN), w ramach projektów Komitetu Badań Naukowych oraz w ramach badań statutowych IPI PAN. Teksty wchodzące w skład Korpusu IPI PAN są dostępne w postaci binarnej, umożliwiającej łatwe i efektywne przeszukiwanie za pomocą dedykowanego oprogramowania o nazwie Poliqarp. [...] Wszystkie zasoby wymienione poniżej, z wyjątkiem Korpusu *Słownika frekwencyjnego*..., dostępne są obecnie na zasadach opisanych w niniejszej umowie licencyjnej, bezpłatnie. [...] 2. wydanie Korpusu IPI PAN (marzec 2006): próbka Korpusu IPI PAN dostępna na stronie http://korpus.pl/; ponad 30 mln. segmentów. Podobnie jak to miało miejsce w wypadku wydania 1., niniejsza wersja sample jest korpusem różnorodnym o następującym składzie: proza współczesna: ponad 10%, proza dawna: prawie 10%, teksty książkowe niebeletrystyczne (głównie naukowe): 10%, prasa: 50%, stenogramy sejmowe i senackie (w tym z komisji śledczej): 15%, ustawy: 5%. Korpus sample jest korpusem różnorodnym, choć być może nie zasługującym na miano reprezentatywnego, zawierającym następujące rodzaje tekstów: proza współczesna: 10%, proza dawna: 10%, nauka: 10%, prasa: 50%,

Języka Polskiego PWN[4]) i dwa korpusy czeskie (ČNK, Korpus DIALOG 0.1[5]).

stenogramy sejmowe i senackie (w tym z komisji śledczej): 15%, ustawy: 5%. Wszystkie teksty, z wyjątkiem prozy dawnej i nielicznych fragmentów prozy współczesnej, pochodzą z ostatnich 15 lat. Proza dawna to przede wszystkim dzieła z końca XIX w. i początku XX w. Jej obecność w korpusie uzasadniona jest obecnością takiej prozy w świadomości zbiorowej Polaków za pośrednictwem lektur szkolnych i ekranizacji. frek.bin.tar.bz2 – kolejna wersja korpusu ***Słownika frekwencyjnego polszczyzny współczesnej*** (Kurcz, Lewicki, Sambor, Szafran, Woronczak 1990, Instytut Języka Polskiego PAN, Kraków). Korpus ten został stworzony w latach 60. ubiegłego stulecia i zawiera pół miliona słów – po 100 tys. słów z: tekstów popularnonaukowych, drobnych wiadomości prasowych, tekstów publicystycznych, prozy artystycznej oraz dramatu artystycznego. Wzbogacona postać tego korpusu, zwana *Wzbogaconym Korpusem Słownika Frekwencyjnego*, dostępna jest na prawach Powszechnej Licencji Publicznej GNU na stronie http://www.mimuw.edu.pl/polszczyzna/. Więcej informacji o kolejnych wersjach korpusu można znaleźć w artykule dostępnym ze strony http://www.mimuw.edu.pl/~jsbien/MO/JwP03/" (www.korpus.pl).

[4] „Wydawnictwo Naukowe PWN przygotowało i udostępniło sieciową wersję Korpusu Języka Polskiego PWN wielkości 40 000 000 słów. Korpus składa się z fragmentów 386 różnych książek, 977 numerów 185 różnych gazet i czasopism, 84 nagranych rozmów, 207 stron internetowych oraz kilkuset ulotek reklamowych. Pełna wersja sieciowa korpusu jest dostępna odpłatnie, a bezpłatnie wersja demonstracyjna wielkości ponad 7,5 miliona słów. [...] Korpus to dowolny zbiór tekstów, w którym czegoś szukamy. O korpusach w tym znaczeniu mówią najczęściej językoznawcy, ale także archiwiści, historycy i informatycy. Korpus tekstów polskich to fragment słownikowej kuchni, czyli autentyczny materiał językowy, na którego podstawie opisujemy znaczenia słów i konstrukcji. Zrównoważenie korpusu jest równie ważne jak jego wielkość. Drobne fragmenty tekstów korpusu są zamieszczane w postaci pojedynczych zdań w słownikach jako przykłady ilustrujące znaczenia. [...] Nasz korpus składa się z tekstów książek, czasopism, druków ulotnych i akcydensowych (np. reklam, instrukcji obsługi, regulaminów, ulotek wyborczych), stron internetowych oraz tekstów mówionych. Teksty książek staramy się pozyskiwać od wydawców w wersji elektronicznej, pytając przy tym o zgodę autorów. Współczesne teksty prasowe przegrywamy z wydań internetowych lub otrzymujemy od redakcji. Starsze teksty prasowe, rzadko wznawiane książki oraz druki ulotne skanujemy. Teksty mówione nagrywamy bezpośrednio (za zgodą mówiących) lub z radia i telewizji, po czym przepisujemy. Stosujemy w nich tradycyjną ortografię (nie alfabet fonetyczny), ale zachowujemy wszystkie powtórzenia i przejęzyczenia. [...] Przygotowując materiał do korpusu internetowego, wybieraliśmy losowo fragmenty książek i czasopism z różnych dziedzin wielkości mniej więcej jednego arkusza (40 000 znaków, czyli około 6 000 wyrazów), założywszy uprzednio strukturę tematyczną. Korpus internetowy dostępny jest w dwóch wersjach: demonstracyjnej i pełnej wersji sieciowej, różniących się od siebie liczbą próbek tekstowych oraz proporcją źródeł" (www.korpus.pwn.pl, www.sjp.pwn.pl).

Dla pełności obrazu właściwym wydaje się wyjaśnienie terminu *korpus*[6]. Dopiero 9 znaczenie tego wyrazu wg *SJP PWN*[7] spełnia nasze wymogi: „teksty, dane itp. zgromadzone ze względu na swą reprezentatywność, stanowiące podstawę do analizy naukowej". Można doprecyzować to znaczenie w interesującym nas sensie lingwistycznym – jest to zespół udokumentowanych dowodów autentycznego

[5] Projekt lingwistyczny Korpus DIALOG to specjalistyczny korpus czeszczyzny mówionej. Prace nad nim rozpoczęto w roku 1997, gdy uruchomiono interdyscyplinarny projekt grantowy na lata 1996–2001 pt. *Dialog w świecie ludzi i maszyn* (*Dialog ve světě lidí a strojů*). Zawiera wypowiedzi publiczne typu dialogowego (wywiad, dyskusja, debata, polemika, talkshow) i obejmuje on zapisy ok. 360 programów dyskusyjnych publicznej i komercyjnych telewizji czeskich (np. *Sedmička, Nedělní partie, Na plovárně s..., Krásný ztráty*). Wielkość korpusu szacowana jest na 2 000 000 wyrazów. Korpus służy do badań nad czeskim językiem mówionym, do opisu stanu jego wersji publicznej oraz do śledzenia jego tendencji rozwojowych. Wykorzystywany jest także do rozwoju teorii wypowiedzi, dialogu i dyskursu. Dla szerokiej publiczności internetowej została udostępniona jego kolejna wersja – Korpus DIALOG 0.1. Korpus ten powstał na bazie Instytutu Języka Czeskiego Akademii Nauk RCz (ÚJČ AN ČR) w ramach projektu grantowego Agencji Grantowej AN RCz (GA AV ČR) pt. *Czeski język mówiony w dyskusyjnych programach telewizyjnych* (*Mluvená čeština v televizních diskusních pořadech 2003–2005*). Projekt ten był realizowany we współpracy z Instytutem Lingwistyki Formalnej i Stosowanej Wydziału Matematyczno-Fizycznego Uniwersytetu Karola w Pradze (ÚFAL M-FF UK). W projekcie tym uczestniczyli doświadczeni badacze współczesnego języka czeskiego: Světla Čmejrková, Lucie Jílková, Petr Kaderka, Jana Klímová, Kamila Mrázková, Zdeňka Svobodová) oraz Nino Peterek (autor rozwiązań technicznych projektu). Korpus działa w oparciu o menadżer obsługi Manatee/Bonito opracowany przez Wydział Informatyki Uniwersytetu Masaryka (FI MU) w Brnie. Zasoby korpusu DIALOG 0.1 obejmują

użycia języka naturalnego, rozległy zespół elektronicznych tekstów celowo zgromadzony jako referencyjne źródło dla naukowej analizy języka. Rozbudowując powyższe znaczenie: k o r p u s j ę z y k o w y to bardzo rozległy kompleks tekstów języka naturalnego, którego powstanie i dalsze używanie możliwe jest za pomocą komputera. Jest to zazwyczaj bardzo bogaty i skomplikowany system tekstów umożliwiający bardzo wydajną metodę analizy językoznawczej nowej generacji. Zastosowanie korpusu jest pewnego rodzaju radykalnym przełomem w lingwistyce, którego pokłosiem stało się powstanie lingwistyki korpusowej. Korpus stanowi kompleks komputerowo zapisanych tekstów – w wypadku języka mówionego jest nim zapis (a nawet transkrypcja) nagrań wypowiedzi – stanowiących bazę do dalszych badań językowych. Do aktywnego korzystania z jego zasobów służy specjalny program wyszukujący. Przy jego użyciu można wyszukiwać wyrazy i konstrukcje wyrazowe w kontekście. Dodatkowo można określić ich frekwencję w korpusie oraz pierwotne źródło tekstowe. W dalszej kolejności możliwa jest dalsza obróbka (analiza) znalezionych haseł, np. porządkowanie alfabetyczne czy też w wypadku niektórych korpusów ekscerpcja według przyjętych kryteriów, np. rodzajów wyrazów. Opracowanie korpusowe języka mówionego polega na stworzeniu dostępnych źródeł referencyjnych tego typu komunikacji werbalnej, co często jest trudne i skomplikowane, gdyż język mówiony – jak polszczyzna tak i język czeski – z filogenetycznego i ontogenetycznego punktu widzenia jest prymarną formą komunikacji językowej i w rzeczywistości (praktyce codziennej) uczestniczy w niej aż w 90%[8]. Tworzenie tego typu wzorcowych zbiorów leksyki i struktur leksykalnych – a w konsekwencji obowiązujących standardów językowych – poprzez zbieranie i obróbkę materiału językowego, udostępnianie „zawsze i wszędzie" oraz wykorzystanie wyników do innych badań lingwistycznych (np. frekwencja jednostek leksykalnych, psycholingwistyka, socjolingwistyka, itp.) i celów pragmalingwistycznych (np. tworzenie i redakcja podręczników języka, rozmówek itp.), staje się działaniem powszechnym także w zakresie języków słowiańskich.

W wielu ośrodkach językoznawczych budowane są różne typy elektronicznych korpusów językowych w zależności od celów badawczych. Z uwagi na ich zakres możemy wyróżnić k o r p u s y o g ó l n e i s p e c j a l n e. Typ ogólny starają się uchwycić język w jak najpełniejszym zakresie i pełni, służy do tworzenia słowników. Typ specjalny obejmuje węższy zakres według jakiegoś przyjętego kryterium; może to być korpus autorski (np. korpus dzieł A. Mickiewicza czy K. Čapka), korpus określonego gatunku lub dzieła literackiego (np. dramatu romantycznego, *Lalki* B. Prusa, *Przygód dzielnego wojaka Szwejka* J. Haška), bądź korpus danego dialektu (np. śląskiego, hanackiego). Z historycznego punktu widzenia tworzone są k o r p u s y s y n c h r o n i c z n e i d i a c h r o n i c z n e – pierwsze dokumentują

zapisy programu dyskusyjnego *Sedmička* (*7 čili Sedm dní*) prywatnej telewizji komercyjnej NOVA z lat 1999–2005. Korpus obejmuje 2 wersje: 1) wersja nieopracowana morfologicznie zawierająca 10 zapisów o wielkości 92 000 wyrazów (opcje: całość, poszczególne programy, wyszukiwanie), 2a) wersja ręcznie opracowana (zrewidowana) morfologicznie zawierająca 5 zapisów o wielkości 45 000 wyrazów, 2b) wersja maszynowo opracowana (zrewidowana) morfologicznie zawierająca 9 zapisów z możliwością rozszerzonego wyszukiwania, odsłuchu przykładów dźwiękowych oraz wizualizację krzywej F0 (opcje: zapis z dźwiękiem, przeszukiwanie zapisu z dźwiękiem). Więcej informacji: www.ujc.cas.cz/ oddeleni/index.php?page=DIALOG.

[6] Korpus – „zestaw tekstów językowych zebrany w celu badania jego systemu lub podsystemu" (*Encyklopedia językoznawstwa ogólnego* 1999, s. 319).

[7] Znaczenie wyrazu *korpus* (od łac. *corpus* 'ciało') podaję za internetowym *Słownikiem języka polskiego PWN*: „1) ciało człowieka lub zwierzęcia oprócz głowy i kończyn; 2) zasadnicza część czegoś; 3) główna część budowli; 4) w architekturze pałacowej: centralna część budynku; 5) w architekturze sakralnej: nawowa część kościoła; 6) główna część, na której oparta jest całość jakiegoś urządzenia, przyrządu itp.; 7) jednostka taktyczna składająca się z kilku dywizji lub brygad; 8) grupa żołnierzy mających taki sam stopień wojskowy; 9) teksty, dane itp. zgromadzone ze względu na swą reprezentatywność, stanowiące podstawę do analizy naukowej" (www.sjp.pwn.pl).

[8] Co ciekawe, chyba najpełniej opracowany i najbardziej reprezentatywny Brytyjski Korpus Narodowy (BNC) posiada największą reprezentację mówionych jednostek językowych, bazując jedynie na 10% tekstów mówionych.

język współczesny, są szeroko używane, a z uwagi na źródła nieocenionym zbiorem informacji o najróżniejszych zjawiskach językowych i pozajęzykowych oraz ich występowaniu i używaniu w naturalnych kontekstach. Korpusy diachroniczne obrazują język starszy, w przeciwieństwie do synchronicznych oparte bywają na wzorcach tekstowych o rozpiętości zazwyczaj 2–5 000 wyrazów, ich tworzenie jest bardzo pracochłonne (elektroniczne skanowanie i ręczne przepisywanie tekstów), stąd ich ilość jest znacznie ograniczona. Aspekty sposobu komunikacji uwzględniają korpusy języka mówionego i korpusy języka pisanego. Z uwagi na pierwotność komunikacji mówionej, redakcja (zestawienie) takich korpusów jest bardzo czaso- i pracochłonne (w pierwszej kolejności zapis i transkrypcja nagrań audio, a następnie opracowanie lingwistyczne tekstów). Korpusy języka pisanego bazują na gotowych tekstach książek, gazet i czasopism najczęściej w zapisie elektronicznym, jednak i te trzeba poddać obróbce formalnej – ujednolicić format, „wyczyścić" z grafiki i ilustracji oraz anotować, czyli opatrzeć notkami (danymi lub symbolami) o charakterze identyfikacyjnym (bibliograficznym), strukturalnym (segmentacja tekstów ciągłych na rozdziały, akapity, zdania i wyrazy) i lingwistycznym (lematyzacja[9], charakterystyka morfologiczna, słowotwórcza, syntaktyczna i ew. semantyczna[10]). Te czynności wykonywane są teraz automatycznie przez specjalne oprogramowanie komputerowe, podobnie jak konkordacja[11] zjawisk i form, czyli występowania danej jednostki korpusowej w wybranym (zadanym)

[9] *Lematyzacja* (od grec. *lémma, lémmatos* 'twierdzenie') to „hasłowanie przy maszynowym przetwarzaniu języka naturalnego. […] W leksykografii przyporządkowanie jednostkom tekstowym nazw jednostek (haseł) opisywanych w słowniku, czyli w istocie napisów służących do wprowadzania ich do słownika. W wypadku najczęstszym hasłowanie polega na przypisywaniu słowu tekstowemu formy podstawowej leksemu opisywanego w słowniku, np. słowu *psa* – hasła *pies* [...]" (*Encyklopedia językoznawstwa ogólnego* 1999, s. 234, 336)

[10] *Tagowanie* to przypisanie tzw. *tagów*, czyli symboli/znaków dodawanych do form wyrazowych, charakteryzujące je pod względem gramatycznym i stylistycznym.

[11] *Konkordacja* (od łac. *concordia* 'zgoda, jedność') jest to „zestawienie wszystkich elementów danego typu występujących w danym tekście lub korpusie. Ele-

przez użytkownika dostatecznym kontekście oraz kombinacja wyrazów. Ostatni podział wyróżnia korpusy jednojęzykowe i wielojęzykowe (paralelne[12]), w których wykorzystywane są specjalne programy zestawiające (parujące) tzw. *leaners' corpora* lub *aligners*, które obydwa zestawy tekstów „układają" obok siebie tak, by zdania, wyrazy i ich połączenia w obu językach korespondowały ze sobą. Takie korpusy mają pierwszorzędne znaczenie w praktyce translacji, gdyż proponują znacznie bogatszą paletę ekwiwalentów tłumaczeniowych wyjściowego wyrazu, frazeologizmu lub zdania, niż dotychczasowe słowniki przekładowe. Niejednokrotnie ich skonfigurowany zasób leksykalny i frazeologiczny stanowi bazę coraz doskonalszych translatorów komputerowych, których sprawność ekwiwalencji osiąga ostatnio nawet 90%.

Na przełomie XX i XXI wieku Uniwersytet Karola w Pradze[13] i Uniwersytet Masaryka w Brnie oraz Instytut Języka Czeskiego AN RCz (ÚJČ AV ČR) stały się bardzo silnymi ośrodkami lingwistyki korpusowej o znaczeniu co najmniej europejskim, o sporym dorobku nie tylko teoretycznym, ale i praktycznym. Dużym osiągnięciem językoznawców czeskich jest liczący sobie już prawie 14 lat elektroniczny (internetowy) Czeski Korpus Narodowy – Český národní korpus (ČNK)[14], będący rozległym grantem akademickim, którego celem jest stworzenie komputerowego korpusu przede wszystkim czeszczyzny

mentami takimi bywają znaki, daty, terminy, zwroty, pojęcia, motywy tematyczne, sentencje, ale najczęściej są nimi słowa podane ze swoimi kontekstami. [...] Obecnie konkordację sporządza się najczęściej na komputerze stosownie do zamówienia badacza – z tekstu lub korpusu zapisanego na nośniku elektronicznym" (*Encyklopedia językoznawstwa ogólnego* 1999, s. 310).

[12] To typ korpusów zestawiających jednakowo treściowe teksty w różnych językach – tekst rodzimy w sąsiedztwie jego obcojęzycznego przekładu (przekładów).

[13] Ojcem projektu i jego duchem sprawczym jest wybitny czeski językoznawca prof. F. Čermák, który „zaraził" pomysłem młodych badaczy i stworzył z nich prężny zespół prowadzący kilka projektów grantowych związanych z ČNK.

[14] Adres internetowy ČNK: www.ucnk.ff.cuni.cz, od października 2007 r. również: www.korpus.cz.

pisanej. Projekt ten stanowi przełom w historii czeskiej lingwistyki, by nie powiedzieć rewolucję w podejściu do badania języka i nawiązuje do najlepszych tradycji czeskiego językoznawstwa (np. Pražský lingvistický kroužek). Opis języka zakłada możliwie na największym zasobie danych językowych – na setkach milionów form wyrazowych, których występowanie i frekwencję może ocenić za pomocą różnych metod matematycznych i statystycznych.

Zapleczem naukowym ČNK stał się Instytut Czeskiego Korpusu Narodowego działający na Wydziale Filozoficznym Uniwersytetu Karola w Pradze (Ústav Českého národního korpusu FF UK).[15] Od swego powstania w roku 1994 zadaniem ÚČNK jest opracowanie i rozbudowanie ČNK oraz działania wspierające, szczególnie w dziedzinie badań i popularyzacji dziedziny lingwistyki korpusowej. Przełomowym osiągnięciem tego znaczącego i zasłużonego centrum lingwistyki komputerowej i korpusowej jest opracowanie ponad stumilionowego korpusu synchronicznego tekstów pisanych SYN2000[16]. W pracach nad ČNK aktywnie uczestniczą również lingwiści i informatycy z prężnego morawskiego ośrodka bohemistycznego – Instytutu Języka Czeskiego Wydziału Filozoficznego oraz Wydziału Informatyki Uniwersytetu Masaryka w Brnie (Ústav českého jazyka FF MU, FI MU). Pracownia brneńska od samego początku aktywnie uczestniczy w czeskich badaniach dot. lingwistyki korpusowej oraz tworzeniu ČNK, specjalizując się redakcji programów komputerowych do automatycznej analizy morfologicznej języka mówionego oraz gromadzi i elektronicznie opracowuje dla potrzeb ČNK specyficznie trudne teksty, np. wypowiedzi mówione i teksty prywatnej korespondencji. Obydwie placówki opracowały dodatkowo korpusy miej-

[15] Od marca 2007 roku ÚČNK ma nową siedzibę, pałac Platýz na Národní třídě 37.

[16] Struktura tego korpusu: teksty publicystyczne – 60%, teksty specjalistyczne – 25%, teksty beletrystyczne – 15%.

skich (regionalnych) odmian języka czeskiego na bazie SYN2000[17]: Pražský mluvený korpus (działa od 2001 roku) a Brnenský mluvený korpus (2002). W placówce praskiej prowadzono także prace nad korpusami paralelnymi InterCorp – na lata 2005–2011 przyjęto projekt grantowy pt. *Český národní korpus a korpusy dalších jazyků*, którego celem jest zbudowanie paralelnych korpusów synchronicznych dla większości języków obcych studiowanych na UK w Pradze (w planach 28 języków), zawsze dla danego języka i czeszczyzny. Projekt ten ma szersze niekomercyjne cele; w oparciu o zgromadzone dane będą prowadzone studia teoretyczne z leksykografii, translatologii, metodyki nauczania języków obcych, opracowane zostaną komputerowe aplikacje do nauki i przekładu języków obcych. W fazie początkowej poszczególne pracownie filologii narodowych w obrębie UK stworzą pod nadzorem i opieką merytoryczną koordynatora programu korpusy narodowe języków obcych[18], które zostaną w fazie późniejszej scalone i udostępnione publicznie na centralnym serwerze projektu.

Pierwotnie struktura Czeskiego Korpusu Językowego obejmowała kilka korpusów synchronicznych:

1) wspomniany już korpus tekstów pisanych **SYN2000**[19] (pojemność 100 mln form wyrazowych);
2) korpus **PUBLIC** (20 mln, 1/5 całości leksyki korpusu SYN2000);

[17] SYN2000 w liczbach: wielkość danych – 2 GB, ilość jednostek tekstowych – 3 303, ilość struktur zdaniowych – 7 639 321, ilość wyrazów tekstowych (*tokens*) – 100 000 704, ilość różnych typów – 1 763 813.

[18] Przy komponowaniu korpusów narodowych w projekcie InterCorp wykorzystane zostaną następujące źródła tekstowe: portale www Unii Europejskiej, dokumenty UE, przepisy prawne UE, projekty Konstytucji Europejskiej i ustaw Traktatów Europejskich (20 języków), SEELRC – The Slavic and East European Language Resource Center, Wikipedia, ACQUIS COMMUNAUTAIRE Multilingual Corpus (multijęzykowy korpus na bazie tekstów legislatywnych UE od roku 1950 do współczesności), Multext-East – „1984" Corpus, databazeUCL.doc (wykaz wolno dostępnych źródeł opracowany przez Instytut Literatury Czeskiej AN RCz).

[19] Na bazie SYN2000 powstał korpus FSC2000 i jego wersja drukowana, będąca słownikiem frekwencyjnym języka czeskiego (F. Čermák, M. Křen, *Frekvenční slovník češtiny*, Praha 2004).

3) korpusy na CD ROM – korpus **SYNEK** (10 mln, 1/10 leksyki), korpus **LITERA** (ok. 3 mln, bazuje tyko na dziełach literackich); korpus **ORWELL** (zasoby na bazie powieści *Rok 1984* G. Orwell'a);
4) korpusy języka mówionego – **PMK** (Pražský mluvený korpus, 800 tys. form wyrazowych) i **BMK** (Brněnský mluvený korpus, 600 tys.).

W październiku 2005 roku ČNK wraz z Czeską Biblioteką Narodową uruchomił dla publiczności DČNK (Diachronní složky ČNK), które są dalej rozbudowywane; zasoby tego korpusu obejmują teksty z ostatnich 700 lat czeskiej literatury (ok. 700 000 form wyrazowych), co roku przybywa ok. 250 000 nowych jednostek. Na bazie DČNK powstał korpus DIAKORP **(**www.ucnk.ff.cuni.cz/diakorp.html**)**, do którego włączono również powstałe do 1989 roku teksty publicystyczne, specjalistyczne oraz artystyczne (do roku 1944). Przełomowy dla tego korpusu może być rok 2008, gdy planowana jest rozszerzona lematyzacja w oparciu o tzw. hiperlemmaty (np. *kůň*), czyli wszystkie wersje graficzne występowania leksemu bez względu na jego różny historycznie zapis (*kóň/kuoň*).

W latach następnych nastąpiły istotne fakty dla bogactwa zasobów i całokształtu działania ČNK, w tym pojawiły się kolejne zaktualizowane i unowocześnione wersje:
– styczeń 2006 – **SYN2005** (100 mln wyrazów tekstowych tzw. *tokens*)[20];
– czerwiec '2006 – **KSK-DOPISY** (Korpus korespondencji prywatnej, zawiera zapisy 2 tys. ręcznie pisanych listów z lat 1990-2004, projekt autorstwa ÚČJ FF MU Brno);
– lipiec 2006 – zakończenie pełnej lematyzację i adnotacji SYN2005;
– listopad 2006 – Český mluvený korpus **ORAL2006** (Czeski korpus języka mówionego, 221 nagrań z lat 2002-2006 o pojemności 1 mln wyrazów);

[20] Korpus SYN2005 (w porównaniu z SYN2000) oparł się na nieco innej strukturze źródeł: beletrystyka – 40%, publicystyka – 33%, literatura specjalistyczna – 27%.

– grudzień 2006 – **SYN2006PUB** (synchroniczny niereprezentatywny korpus publicystyki pisanej o pojemności 300 mln tokens),
– styczeń 2007 – **Bonito2** (www.ucnk.ff.cuni.cz/corpora; nowoczesna i wielofunkcyjna przeglądarka i wyszukiwarka);
– grudzień 2007 – dodano **Inverse Text Sort** (program do wstecznego segregowania zasobów);
– styczeń 2008 – pojemność wszystkich zintegrowanych korpusów w ramach ČNK wyniosła 500 mln form wyrazowych, najnowsza lematyzacja oraz adnotacja morfologiczna (tagowanie).

Współczesną strukturę ČNK ilustruje tabela 1.

Tabela 1. Struktura Czeskiego Korpusu Narodowego (źródło: www.ucnk.ff.cuni.cz)

<table>
<tr><th colspan="3">ČESKÝ NÁRODNÍ KORPUS</th></tr>
<tr><th colspan="2">Część synchroniczna</th><th>Cześć diachroniczna</th></tr>
<tr><td colspan="2">Bank synchronicznego języka czeskiego
zawiera poprawione i skonwertowane współczesne czeskie teksty oraz przeglądarkę/wyszukiwarkę korpusową</td><td>Bank diachronicznego języka czeskiego
zawiera teksty staroczeskie z trzech grup; teksty transkrybowane (2 mln wyrazów tekstowych), teksty transliterowane (ok. 100 tys.) i teksty gwarowe (ok. 200 tys.)</td></tr>
<tr><th>Korpusy pisane</th><th>Korpusy mówione</th><td>Korpus diachroniczny</td></tr>
<tr><td>SYN2006PUB
SYN2005
SYN2000
FSC2000
KSK-DOPISY
SYNEK
LITERA
ORWELL</td><td rowspan="3">Pražský mluvený korpus
Brněnský mluvený korpus
ORAL2006</td><td rowspan="3">Obejmuje wybór tekstów staroczeskich od pierwszych zachowanych zabytków do lat ujętych w korpusie synchronicznym
DIAKORP</td></tr>
<tr><th>Korpusy paralelne</th></tr>
<tr><td>Projekt InterCorp</td></tr>
</table>

ČNK oferuje użytkownikom szerokie zasoby, a w ich obrębie dane szczegółowe:

1) typowe (wskazanie: centralne czy marginalne),
2) aktualne (synchroniczne i aktualne),
3) nieselektywne (niefiltrowane według jakiegoś klucza lub subiektywnie),
4) obiektywne i realistyczne (źródła udokumentowane, rzeczywiście zapisane),
5) dostateczne (wystarczająco rozległe do poznania i opisu danego zjawiska).

Posługując się korpusem możemy wyszukać:

1) konkretną formę wyrazową (wyraz tekstowy z kontekstem);
2) jednostki wielowyrazowe (np. wyrażenia przyimkowe, frazelogizmy);
3) leksem lub hasło kluczowe (tzw. *Lemmat*);
4) części mowy (przymiotnik, symbol/tag: *adjektivum* = „A.*”),
5) kombinację części mowy i lemmatu.

[21] Najczęściej wyświetlane jednostki leksykalne ČNK (SYN2000): formy wyrazowe: *a, se, v, na, je*; rzeczownik: *rok*; przymiotnik: *český*, przysłówek: *už*; imiona męskie: *Jan, Jiří, Václav, Petr, Josef*; imiona żeńskie: *Marie, Jana, Eva, Anna, Hana*; czeskie miasta: *Praha, Brno, Plzeň, Ostrava, Prostějov, Olomouc*; kraje: *ČR, USA, Německo, Rusko, Slovensko, Francie, Polsko, Itálie*; miasta zagraniczne: *Moskva, Paříž, New York, Bratislava, Londýn, Washington, Vídeň, Berlín, Brusel, Řím*; godzina: *20.00*. Nietradycyjna kolejność występowania jednostek wg kryterium frekwencyjnego: liczby: *1, 2, 3, 6, 4, 5, 0, 7, 10, 8, 9*; kontynenty: *Evropa, Amerika, Afrika, Asie, Austrálie, Antarktida*; dni tygodnia: *sobota, neděle, pátek, pondělí, středa, úterý, čtvrtek*; miesiące: *září, leden, květen, listopad, červen, říjen, duben, březen, srpen, prosinec, červenec, únor*; lata: *1994, 1995, 1996, 1997, 1993, 1992, 1998,1991, 1990, 1989, 2000, 1999*; partie polityczne (skróty): *ODS, ČSSD, ODA, KDU, KSČM, HZDS, KSČ*; tytuły naukowe i zawodowe (skróty): *Ing., MUDr., JUDr., PhDr., Mgr., RNDr., DrSc., MVDr., ThDr., PaeDr., RSDr., Bc., RCDr*. (www. ucnk.ff.cuni.cz/korpus/korpus.html).

[22] Najdłuższe czeskie wyrazy: *luftwafehilfefrauenfunkršýlerin, riptopolymernerelativistických, ethylendioxyfenylisopropylamin, českomoravskoslezskocikánského, ukrajinskovietnamskolaoskočeský, technologickoekonomickofinanční, nikotinamidadenindinukleotidfosfát, ultrasuperkontramultiextraunikátní, cyklopentanoperhydrofenantrenový, glyceraldehydfosfátdehydrogenázou, komunistickosociálněmokratického, hypotalamohypofyzoadrenokortikální, francouzočeštinolaosoruštinoindočínštinou*. (www. ucnk.ff.cuni.cz/korpus/korpus.html)

Oprócz charakterystyki statystycznej i frekwencyjnej (częstotliwość występowania jednostki wyrazowej w korpusie czy języku, najpopularniejsze[21] oraz najdłuższe[22] wyrazy czeskie), jest to wspaniałe narzędzie dające możliwości weryfikacyjne nie tylko specjalistom (językoznawcom-bohemistom), ale także miłośnikom starannej czeszczyzny, językowym purystom oraz szerokiemu gronu użytkowników, sprzyjając w ten sposób poprawności językowej, spełniając wymogi uzusu semantycznego. Dzięki ČNK możemy badać i opisywać łączliwość (tzw. kolokację) poszczególnych jednostek leksykalnych z innymi, weryfikować jednostki frazeologiczne oraz walencję wyrazową czy też rekcję czasowników[23]. Jako źródło elektroniczne „on line” może reagować natychmiast na pojawienie się „nowinek” w leksyce współczesnej czeszczyzny, których nawet najnowsze drukowane słowniki języka nie uwzględniają (np. wyrazy pochodne); sprawdzić, który typ deklinacyjny dany leksem reprezentuje, czy dany wyraz (neologizm) nie ma odmiany mieszanej[24]. Użytkownicy współczesnego języka czeskiego – rodowici Czesi i obcokrajowcy – dzięki ČNK mogą poznawać, wzbogacać i porównywać ojczysty i obcy zasób leksykalny (np. *briefing/brifink, football/fotbal*); skonfrontować warianty leksemów (np. *alespoň/aspoň, ačkoliv/ačkoli*) i ich stylistyczny ładunek. W korpusie bez trudności można poznać bogactwo cze-

[23] Dla Polaków uczących się języka czeskiego rekcja czasowników czeskich o brzmieniu podobnym do polskich (często aproksymaty) może sprawić wiele niespodzianek.

[24] Tzw. „chwiejność” rodzaju gramatycznego, np. czeski rzeczownik *rukojmí* i jego „niejasny” paradygmat odmiany i łączliwość np. z liczebnikami.

[25] Przykładowe ćwiczenia zainspirowane ČNK, do których dołączony jest klucz:

1. Určete vynechané slovo:

Válka o <.....> země však již trvá čtyřicet osm let.
Člověk potřebuje k životu dvě věci: <.....> pravdy a ždibec salámu.

skiej synonimii (np. określenia wartościujące). Dla bohemistów-dydaktyków języka czeskiego (zwłaszcza cudzoziemców) jest to świetny zasób/źródło ćwiczeń nie tylko leksykalnych[25] – można za jego pomocą, poprzez celowe opuszczenie słów uczyć i ćwiczyć poprawność pozycji danego wyrazu tekstowego w kontekście czy też szeregu syntagmatycznym. I chyba ostatnie praktyczne zastosowanie – dane ČNK są nieocenione przy redagowaniu słowników, logicznych spisów/wykazów informacji oraz obsługi translacyjnej osób niesłyszących, przy której należy optymalnie ograniczyć ilość środków komunikacyjnych do tych najniezbędniejszych.

Jak wynika z powyższego funkcjonalność i wielowymiarowość ČNK jest coraz bardziej doceniana przez bohemistów (w tym nauczycieli języka i tłumaczy), językoznawców i specjalistów od komunikacji społecznej (dziennikarzy, twórców reklam), ale także przez ogół użytkowników współczesnego języka czeskiego. Jest to źródło coraz bardziej popularne i opiniotwórcze, o czym świadczy rosnąca z mie-

---

Přistanu o <.....> dál, na konci pšeničného pole.. [...]

5. Určete vynechané slovo a všechny jeho tvary, ve kterých se v konkordančních řádcích vyskytlo:

   studijní dokumentace na všech fakultách s <.....> lékařské poskytovala totiž komisím možn
   Trval na své žádosti a byl přijat až na <.....> ministra vnitra . Od roku 1988 na Techn
   ekonomiku celého státu i všech občanů bez <.....>. Lze proto právem očekávat , že zejmén

6. Určete, v kterém pádě jsou podstatná jména v uvedených kontextech použita a tvary doplňte:

   sociální pojištění atd. atd.) bylo svěřeno <odborník...> a prodebatováno, takže program
   publikování před možností podrobit svůj talent <zkouš.....> opravdovosti, v níž se literární
   smrti jeho otce Hudžra, zasvětil svůj život <pomst.....> a úsilí o obnovení moci svého rodu
   to říkala dvakrát. Chcete snad nyní <otc.....> něco vytýkat ? A nebo snad... [...]

7. Doplňte složená slova s první částí *polo-* v náležitém tvaru:

   mohou být výsledky jeho cesty na Malajský <polo.....> oboustranně příjemným překva
   změní v mrznoucí déšť. Ve čtvrtek bude <polo.....> až oblačno a ojediněle přeháňky. Noční
   se jedno setkání ODA s voliči konalo před <polo.....> sálkem Městské knihovny v Praze...

8. Doplňte složená slova se základem vláda:

   sovětský politik postupně chápat, že <.....vláda> není pro kontraproduktivní sovětský systém
   pomine jako špatné zboží. Jenom cáři, <.....vláda> a takové sloty potrvají věčně věkův jak

siąca na miesiąc rzesza osób odwiedzających portal ČNK, wyszukujących hasła i konsultujących swe wątpliwości. W miarę wzbogacania zasobów Korpusu oraz możliwości obliczeniowych użytkowanego przez projekt sprzętu i oprogramowania, jesteśmy świadkami powstawania i krzepnięcia oraz dalszego optymalizowania funkcji i możliwości tego nowoczesnego i wszechobecnego dzięki Internetowi narzędzia lingwistycznego, którego dalsze dziedziny zastosowań w naszym życiu – nie tylko naukowym i akademickim – są wszechstronne i nieodgadnione. Już niebawem przekonamy się o tym.

Literatura

Blatná R., 1997a, *Jazyková databanka neboli korpus*, *Vesmír*, č. 12, s. 670–671.

Blatná R., 1997b, *Korpus – uskutečněný sen*, *Tvar*, č. 9, Praha, s. 17.

Čermák F. a kol., 2007, *Frekvenční slovník mluvené češtiny*, Praha.

Čermák F., Blatná R., 2006, *Korpusová lingvistika: Stav a modelové přístupy*, Praha.

Čermák F., Blatná R., 2005, *Jak využívat Český národní korpus*, Praha.

Čermák F., Klímová J., Petkevič V., 2000, *Studie z korpusové lingvistiky*, Praha.

Čermák F., Kubíček P., 1997, *Jazykový korpus a škola*, „Český jazyk a literatura" XLVIII, č. 3–4, Praha, s. 84–92.

Čermák F., Křen M., 2004, *Frekvenční slovník češtiny*, Praha.

Čmejrková S., Jílková L., Kaderka P., 2004, *Mluvená čeština v televizních debatách: korpus DIALOG*, „Slovo a slovesnost" LXV, Praha, s. 243––269.

Encyklopedia językoznawstwa ogólnego, 1999, red. K. Polański, Ossolineum.

Kocek J., Kopřivová M., Kučera K., 2000, *Český národní korpus – úvod a příručka uživatele*, Praha.

Sborník Asociace učitelů češtiny jako cizího jazyka (AUČCJ) 2003–2005, Praha 2005, s. 11–16.

Šulc M., 1999, *Korpusová lingvistika. První vstup*, Praha.

www.korpus.pl [stan z 15.08.2008 r.]

www.ujc.dialogy.cz [stan z 15.08.2008 r.]

www.ucnk.ff.cuni.cz [stan z 15.08.2008 r.]

www.www.korpus.cz/intercorp [stan z 15.08.2008 r.]

www.korpus.pwn.pl [stan z 15.08.2008 r.]
www.sjp.pwn.pl [stan z 15.08.2008 r.]
www.ucnk.ff.cuni.cz/diakorp.html [stan z 15.08.2008 r.]