Prehľad textových korpusov slovanských jazykov

Alexander Horák
Oddelenie Slovenského národného korpusu, Jazykovedný ústav Ľ. Štúra SAV, Bratislava

Príspevok bol prednesený na XII. kolokviu mladých jazykovedcov v Modre a bude publikovaný v zborníku Varia z tejto konferencie.

Abstrakt

Cieľom príspevku je poskytnutie prehľadu o dostupnosti existujúcich textových korpusoch v niektorých slovanských jazykoch, ich porovnanie a načrtnutie možnosti a spôsobu ich využívania v lingvistickom výskume.

1 Korpusy a korpusová lingvistika

O poslednom desaťročí 20. storočia sa často hovorí ako o desaťročí korpusovej, prípadne počítačovej lingvistiky. Zrejme celkom oprávnene. Veľké textové súbory – korpusy ako prirodzené prejavy jazykovej performancie boli predmetom lingvistickej analýzy hádam od jej začiatkov, avšak táto analýza nadobudla nový rozmer rozvojom informačných technológií a počítačových sietí a ich vstupom a masívnym rozšírením do humanitných vied práve v 90. rokoch 20. storočia. Okrem prakticky nevyčerpateľnej možnosti byť archivačným médiom, počítače poskytli aj možnosť automatizovanej lingvistickej analýzy textov na rozličných úrovniach, ktorá by sa nevmestila do ľudských kapacít. A navyše, s istým nádychom obraznosti by sa dalo povedať, že korpusy sa stali zdokumentovaným ľudským poznaním v istom období jeho vývinu. Dôkladné opísanie významu a vývinu korpusovej lingvistiky by však presahovalo rámce tohto článku a koniec koncov, nazdávame sa, že tak bolo učinené už mnohokrát doteraz na dostatočne vysokej úrovni [1]. Ďalej sa preto zameriam na popis súčasného stavu v oblasti, pričom uvediem niekoľko pojmov, nevyhnutných pri následnom kvalitatívnom porovnávaní korpusov slovanských jazykov.

1.1 Elektronické textové databázy

Korpus textov nejakého jazyka sa zvyčajne definuje akovnútorne štrukturovaný, unifikovaný, oindexovaný a ucelený rozsiahly súbor elektronicky uložených a spracovávaných jazykových dát väčšinou v textovej podobe. Líši sa tým od iných textových databáz ako napríklad elektronický archív, čo je sklad elektronicky čitateľných textov, ktoré nie sú spojené ani spracované (napr. Oxfordský textový archív) alebo elektronickej knižnice (textotéky) – zbierky elektronických textov spojených podľa istého obsahového kritéria.

1.2 Typy korpusov

Členenie korpusov je možné podľa viacerých kritérií:

- podľa počtu obsiahnutých jazykov na:
  - jednojazyčné, ktoré sú tvorené väčšinou ako reprezentatívne korpusy konkrétneho národného jazyka
  - viacjazyčné (paralelné) využiteľné napr. pri strojovom preklade, prípadne pri zostavovaní prekladových slovníkov
- podľa časového záberu textov v korpuse na :
  - synchrónne, ktoré zachytávajú súčasný jazyk
  - diachrónne, ktoré sú dokumentom istého obdobia jazykového vývinu
- podľa východiskovej formy textov v korpuse na:
  - korpusy písaného textu, ktorých je v súčasnosti prevažná väčšina
  - korpusy hovoreného textu, ktoré sa pre technickú náročnosť budovania vyskytujú pomerne zriedkavo, avšak z hľadiska lingvistickej analýzy sú o to vzácnejšie pre spontánny charakter hovorených jazykových prejavov
- podľa funkcie/špecializácie na:

- všeobecné/základné (reprezentatívne, vyvážené)
- špecializované na základe istého lingvistického kľúča (napr. korpus textov nejakého autora, prípadne obdobia, regiónu, štýlu)

- trénovacie korpusy sú osobitným druhom korpusov, použiteľným na nácvik programov, ktoré vykonávajú automatickú lingvistickú analýzu zväčša na úrovni morfológie, prípadne syntaxe. Trénovacie korpusy sú na tieto účely manuálne lingvisticky označené – k jazykovým jednotkám je pridaná ich lingvistická interpretácia (napr. gramatická kategória, vetnočlenská funkcia).

1. 3. Základné vlastnosti korpusu

Pri porovnávaní korpusov je relevantných niekoľko parametrov, ktoré sú zväčša aj kvantifikovateľné. Jednou z podstatných vlastností korpusu je jeho VEĽKOSŤ. Podľa veľkosti, akú korpusy v priebehu času mohli nadobúdať sa dokonca vyčleňuje niekoľko generácií korpusov:

1. generácia korpusov (od 50. rokov do r. 1975)**:** obsahovala od 500 tisíc do 1 miióna textových slov. Sem sa zaraďuje prvý moderný korpus v lingvistike – Brownov korpus (Brown Corpus), ale aj ďalšie ako napríklad, korpus Lancaster-Oslo/Bergen (LOB Corpus), korpus Prehľad hovorenej angličtiny (Survey of Spoken English – SSE), korpus London-Lund (LLC), korpus Nijmegen a ďalšie.
2. generácia korpusov (do r. 1985). Korpusy tejto generácie obsahovali od jedného do 20 miliónov slovných tvarov. V tomto období vymedzenom zhruba od polovice 80. rokov vznikajú početné korpusy, najmä angličtiny, ktoré sa špecializujú na rozličné regionálne varianty angličtiny, hovorený jazyk, jazyk špecifických prostredí atď.
3. generácia korpusov. Vznik korpusov 3. generácie sa datuje od prelomu 80. a 90. rokov 20. storočia, kedy sa začalo budovanie prakticky väčšiny národných korpusov európskych jazykov (napr. Britský národný korpus, Bank of English, Americký národný korpus, Korpus písanej taliančiny, Írsky národný korpus, Český národný korpus, atď.). Ich veľkosť sa rádovo pohybuje v stovkách miliónov, niekedy až miliardách slovných tvarov (napr. Mannheimský korpus nemčiny)

S veľkosťou korpusov súvisí aj ich REPREZENTATÍVNOSŤ, ktorá sa nechápe v zmysle normy, kodifikácie a spisovnosti, ale skôr v zmysle reprezentovania čo najširšieho záberu rozličných jazykových variet, typov a žánrov textov, autorov textu, obdobia vzniku alebo publikovania textov, druhovej alebo štýlovej príslušnosti textov (vedecké, publicistické, umelecké atď.). V súčasnosti za základné atribúty reprezentatívneho národného jazykového korpusu pokladajú:

- veľkosť minimálne 100 miliónov textových slov
- vyváženosť podľa zastúpenia jednotlivých štýlov, žánrov, typov, autorov textov
- lingvistická anotácia, ktorá sa realizuje v dvoch rovinách:
  1. externej, ktorá zachytáva logickú štruktúru textu (kapitoly, odsek, vety, nadpisy), bibliografické údaje (autor, rok vydania, prekladateľ a pod.) a štýlovo-žánrové charakteristiky textu.
  2. internej, ktorá reprezentuje lingvistickú interpretáciu jazykových jednotiek v korpuse na rozličných rovinách (fonetickej, morfologickej, syntaktickej, sémantickej, diskurzovej).

2 Korpusy súčasných slovanských jazykov

2.1 Český národný korpus

Český národný korpus (ČNK) sa často uvádza ako príklad korpusu porovnateľného so západnými korpusmi svojou štruktúrou, metódou jazykovedného a počítačového spracovania a tiež aj veľkosťou jeho synchrónnej reprezentatívnej časti (viac ako 100 miliónov slov). ČNK je dielom spolupráce Karlovej Univerzity (Ústav teoretickej a komputačnej lingvistiky, Ústav bohemistických štúdií, Katedra českého jazyka Filozofickej fakulty, Ústav formálnej a aplikovanej jazykovedy Matematicko-fyzikálnej fakulty), Českej technickej univerzity v Prahe, Masarykovej Univerzity v Brne (Katedra českého jazyka Filozofickej fakulty, Fakulta informačných technológií) a Českej akadémie vied. ČNK je v skutočnosti

súhrnným názvom pre viacero korpusov – v zásade je rozdelený na dve časti: synchrónnu a diachrónnu. Každá z týchto častí sa potom člení na ďalšie dve: archív a banku podľa formátu v akom sa v nich texty nachádzajú. V archíve sú texty v pôvodných formátoch, v ktorých sa nadobudli (napríklad Microsoft Word, QuarkXpress atď.), kým v banke sú už skonvertované do formátu SGML[1] a pripravené na vyhľadávanie korpusovým manažérom. Zloženie synchrónnej časti na úrovni banky sa ďalej rozčleňuje na jednotlivé korpusy:

- Databázy a slovníky (DB): okrem vlastných korpusov obsahuje ČNK pomocné databázy – napr. Evidence – databáza všetkých korpusových textov, alebo slovníky v elektronickej podobe.
- SYN2000 je reprezentatívny, vyvážený korpus písanej súčasnej češtiny, obsahujúci 100 miliónov slov.
- PUBLIC je korpus prístupný verejne na Internete, obsahuje 20 miliónov slov. Je vytvorený zo SYN2000 a vyvážený rovnako.
- ORAL je korpus hovoreného jazyka, ktorého hlavnú časť zatiaľ tvorí tzv. Pražský mluvený korpus (PMK).
- DIAL – je označenie pre plánovaný nárečový korpus.

Podobne je aj diachrónna časť na úrovni banky rozdelená na viacero korpusov:

- Banka diachronní češtiny (Banka ČNKDIA) sa skladá z banky transkribovaných textov (asi 2 milióny textových slov), banky transliterovaných textov (asi 100 tisíc textových slov) a banky nárečových textov (asi 200 tisíc textových slov).
- Databázy a slovníky (DB) pre diachrónnu časť obsahujú okrem evidenčnej databázy napr. prekladový slovník starších českých slov. Vyhľadávanie umožňuje iný vyhľadávací program, ako v synchrónnych textoch.
- DIAKORP (Diachronní korpus) obsahuje výber staročeských textov od prvých zachovaných záznamov po dobu pokrytú synchrónnym korpusom.

V priebehu tvorby ČNK existovali viaceré programy na jeho analýzu: zo začiatku sa používal program Stuttgartskej univerzity CQP (Corpus query processor), neskôr sa vyvinuli aj korpusové manažéry určené špeciálne preň, napríklad CQM (Corpus query manager). V súčasnosti ku ČNK používatelia pristupujú pomocou programu GCQP, čo je grafické rozhranie pre korpusový manažér CQP. CQP umožňuje vyhľadávanie postupnosti niekoľkých slov, vyhľadávanie podľa lemy a morfologických značiek, zobrazenie bibliografických údajov pri jednotlivých konkordanciách, štatistické funkcie, vytváranie zložitejších požiadavok na vyhľadávanie používaním regulárnych výrazov a ďalšie nastaviteľné funkcie. Pri menšej časti ČNK, vydanej na CD nosičoch s názvom SYNEK, ktorá je vlastne zmenšeninou SYN2000 (v zmysle zachovania reprezentatívnosti textov) sa používa korpusový manažér Manatee s grafickým rozhraním Bonito. Je plánovaný postupný prechod celého ČNK na tento systém.

2.2 Pražský závislostný korpus

Projektom, ktorý do istej miery vyšiel z ČNK je Pražský závislostný korpus (PZK). Realizuje sa na Ústave formálnej a aplikovanej lingvistiky a Centre komputačnej lingvistiky Matematicko-fyzikálnej fakulty UK v Prahe. Jeho cieľom je anotácia menšieho množstva textov na viacerých jazykových úrovniach, predovšetkým na syntaktickej. Koncepcia anotácie vychádza z teórie funkčného generatívneho opisu, ktorej autorom je Petr Sgall. V súčasnosti korpus pozostáva zhruba z 1,4 milióna anotovaných textových slov na úrovni morfológie a povrchovej syntaxe, v roku 2004 je plánované ukončenie anotácie 1 milióna slov na hĺbkovo-syntaktickej (tektogramatickej) rovine. Na prácu s korpusom sa používajú špecializované nástroje vysokej kvality: manuálna syntaktická anotácia sa vykonáva programom TrEd a vyhľadávanie programom Netgraph.

2.3 Korpus frekvenčného slovníka poľštiny

Najstarším textovým korpusom poľštiny v modernom zmysle je korpus, ktorý bol výskumným

[1] Formát SGML (Standard Generalized Markup Language) vznikol ako medzinárodný štandard pre achiváciu a reprezentáciu jazykových dát v elektronickej podobe. V súčasnosti sa na tieto účely čoraz viac využíva formát XML (eXtensible Markup Language), ktorý je jeho podmnožinou.

materiálom pre Frekvenčný slovník súčasnej poľštiny (*Słownik frekwencyjny polszczyzny współczesnej*) vydaný v r. 1990. Tento korpus obsahuje texty zo šesťdesiatych rokov (od r. 1963 do r. 1967), ktoré sú vyvážené podľa funkčných štýlov v poľštine – vedecko-populárneho, publicistického, umeleckého (próza a dráma). Celkový objem korpusu dosahuje počet 500 000 slovných tvarov. Korpus je obohatený o morfologickú anotáciu, ktorá bola vykonaná ručne. V anotácii sa rozlišuje deväť slovných druhov: substantíva, adjektíva (do ktorých sú zahrnuté aj deverbatívne deriváty a radové číslovky), číslovky, zámená, slovesá, predložky, príslovky, spojky, častice. Tieto sú v texte označené tagom, ktorý má podobu číselnej značky. V notácii značiek sa na prvej pozícii nachádza značka slovného druhu a na ďalších hodnoty príslušnej morfologickej kategórie (napr. pre substantíva pád, číslo, pre slovesá čas, spôsob, syntetickosť/analytickosť, zvratnosť). Dôležitou vlastnosťou korpusu je, že čiastočne rieši problém tokenizácie viacslovných pomenovaní ako napr. frazeologizované predložkové spojenia (*na razie, w lewo, po polsku*), spojenia *co* a *jak* s adjektívami a príslovkami (*jak najwyżej, jak najbardziej, co najmniej*), zvratné príčastia a slovesné podstatné mená (*skarżący się, zastanowienie się*), cudzie priezviská s *von, de* (*von Beethoven, de Gaulle*) a iné viacslovné jednotky (*mimo że, jak gdyby*). Korpus je zapísaný v podobe holých textových súborov, čo si nevyžaduje špecializované programové nástroje na jeho prehľadávanie.

Korpus frekvenčného slovníka poľštiny sa považuje za doteraz najlepšie spracovaný textový korpus poľštiny napriek tomu, že sa mu niekedy vytýka jeho zastaranosť [3]. Tá sa však týka skôr lexikálneho aspektu analýzy, keďže funkčné štýly v poľštine prešli od šesťdesiatych rokov značnou premenou (najmä publicistický).

2.4 Korpus PELCRA

PELCRA je širokozvetveným poľsko-anglickým programom, ktorý má ambíciu okrem reprezentatívneho Poľského národného korpusu vypracovať porovnávacie poľsko–anglické korpusy, ktoré by boli určené pre poľských záujemcov o štúdium angličtiny alebo prekladateľov. Projekt sa začal v roku 1997 a podieľajú sa na ňom Katedra anglického jazyka Lodžskej univerzity a Oddelenie jazykovedy a súčasného anglického jazyka Univerzity v Lancasteri, ktorá, ako je známe, budovala v spolupráci s Oxfordskou univerzitou aj Britský národný korpus (BNC). Preto sa na WWW stranách tohoto projektu vyslovene deklaruje cieľ zostaviť korpus, ktorý by veľkosťou a štruktúrou plne zodpovedal Britskému národnému korpusu. Má zahrňovať písané a hovorené texty, slovnodruhovo označené a anotované podľa odporúčaní konzorcia TEI[2]. Z plánovaného počtu 130 miliónov slov je v súčasnosti spracovaná 30 miliónová čiastka, v ktorej sa zachovávajú proporcie BNC. Jej sprístupnenie na internete je avizované do „niedalekej przyszłości“ (budúcnosti). Verejnosti prístupný je však Uczniowski Korpus Języka Angielskiego a Polski Multimedialny Korpus Konwersacyjny, ktorý si možno aj vypočuť vo formáte mp3. Kým prvý, 500 tisícový, je zameraný na poukázanie syntaktických a lexikálnych rozdielov medzi angličtinou a poľštinou so zvláštnym dôrazom na ukázanie častých chýb Poliakov vo výučbe angličtiny, v druhom ide o zachytenie autentického hovoreného jazyka. Autenticitu vnáša fakt, že rozhovory boli nahrávané bez vedomia hovoriacich, ktorí boli vyberaní, so zámerom obsiahnuť čo najširšie oblasti jazyka (rovnaké zastúpenie oboch pohlaví, rozličný sociálny pôvod a príslušnosť, vzdelanie)

2.5 Korpus PWN

Je skôr komerčným projektom, ktorý ma charakter materiálovej základne pre slovníky Poľského vedeckého vydavateľstva (Polskie Wydawnictwo Naukowe). Je to takisto referenčný korpus predovšetkým súčasnej poľštiny, t.j. poľštiny 20. storočia (od r. 1918) a zvlášť jeho posledného desaťročia – texty z tohoto obdobia tvoria polovicu korpusu. Jeho celkový objem je okolo 50 miliónov slov, z toho približne dvojmiliónová čiastka je prístupná na internete. Tvorcovia tohto korpusu sa okrem všeobecných zásad zostavovania korpusu riadili aj kritériom špecificky poľskej „tradície kultúrnej autority spisovateľa“ ako meradla spisovnosti.

[2] TEI je skratka pre Text Encoding Initiative, medzinárodné konzorcium, ktoré vydáva odporúčania o spôsobe značkovania a archivácii textov dodatočnými informáciami.

Štruktúra korpusu PWN:

Beletristika (zahrnutá aj poézia)	19%
„Nebeletristika" (vedecká literatúra, príručky, albumy, spomienky, rozhovory)	27%
Publicistika (denná tlač, časopisy)	47%
Hovorené texty	7%

Tematická štruktúra publicistickej literatúry a „nebeletristiky":

filozofia, náboženstvo	6%
história, geografia	9%
literárna veda, jazykoveda, eseje	6%
prírodné a matematické vedy	8%
politika, ekonómia	23%
spoločenské vedy	11%
aplikované vedy	12%
umenie	5%
rekreácia	7%
denná tlač	13%

Časová štruktúra textov:

1918-1945	10%
1944-1970	13%
1970-1989	17%
1990-2000	60%

Tento korpus má uplatnenie hlavne v lexikografii - napríklad pri výskume kontextových významov lexikálnej jednotky, kde je potrebné odlíšenie synonymných významov, nové definície významov a významových odtienkov, pri mapovaní novšej slovnej zásoby a pod. Konkrétnym lexikografickým dielom, vzniknutým na základe tohto korpusu je *Inny słownik języka polskiego*, ktorý vyšiel v roku 2000 v redakcii Mirosława Bańku. Korpus je možné najjednoduchšími metódami prehľadávať pomocou grafického rozhrania, ktoré je na WWW strane vydavateľstva PWN.

2.6 Korpus IPI PAN

Inštitút základov informatiky Poľskej akadémie vied (Instytut Podstaw Informatyki Polskiej Akademii Nauk) buduje ďalší korpus poľštiny, ktorý mal na začiatku len status trénovacieho korpusu pre vývoj programov na automatickú lingvistickú analýzu. V súčasnosti prešiel do polohy osobitného grantu, majúceho za cieľ vytvorenie rozsiahleho anotovaného reprezentatívneho korpusu poľštiny. Korpus pravdepodobne obsahuje viac ako 13,4 milióna slov, taký je však jeho stav, ktorý sa uvádza na WWW strane tohto projektu[3]. Obsahuje publicistické a právne texty, novšiu a klasickú poľskú prózu (Konopnicka, Sienkiewicz, Witkacy), prepisy telefonických rozhovorov a dokonca aj texty Starej a Novej zmluvy. Veľkými nevýhodami verejne prístupnej časti korpusu sú, že nie je vyvážený, anotovaný, dokonca ignoruje poľskú diakritiku (napr. forma pisze môže byť interpretovaná aj ako pisze – 3.os.sg aj ako piszę – 1.os.sg.) a v niektorých prípadoch sa vyskytujú skenovacie chyby (o namiesto s). Viaceré publikácie [4] [2] [7] nasvedčujú tomu, že po zrealizovaní projektu to bude korpus, porovnateľný svojimi parametrami s lepšími korpusmi vo svete.

[3] http://dach.ipipan.waw.pl/CORPUS

2.7 Korpus FIDA

V Slovinsku sa za krátky čas (od polovice r. 1997) podarilo vybudovať korpus FIDA, ktorého hlavnými charakteristikami sú referenčnosť, jednojazyčnosť, synchrónnosť a východiskovo písaná forma textov. Korpus FIDA je výsledkom spolupráce akademických a „priemyselných" partnerov: Filozofickej fakulty Ljubljanskej Univerzity, Inštitútu Jožef Štefan, vydavateľstva DZS (Državna založba Slovenije) a podniku Amebis. V súčasnosti korpus obsahuje okolo 100 miliónov slovných tvarov a jeho prehľadávaniu slúži softvér vyvinutý v domácom prostredí – program ASP32 (Amebisovo skladišče podatkov). Ten sa používa aj pri vyhľadávaní v elektronických verziách slovníkov DZS a tiež v internetovom rozhraní aj na WWW stránkach FIDA[4]. Korpus je lematizovaný a anotovaný morfologicky a aj bibliograficky.

2.8 Chorvátsky národný korpus

Od konca šesťdesiatych rokov existovali v Chorvátsku komparatívne projekty chorvátčiny a angličtiny, ktoré sa zakladali Brownovom korpuse. Neskôr, v priebehu sedemdesiatych rokov, sa pozornosť sústredila na vytvorenie korpusov textov staršej chorvátskej literatúry, ktorá vyvrcholila v známom tzv. Mogušovom korpuse chorvátskeho jazyka, na vtedajšiu dobu ojedinelým projektom v rámci jazykovednej slavistiky: Mogušov korpus mal ambíciu zahŕňať až milión slovných tvarov z textov staršej aj súčasnej chorvátskej literatúry. Na vtedajšiu dobu (1975) to bol parameter porovnateľný s vtedajším stavom korpusov v britskej jazykovede (napr. Brownov korpus). Mogušov korpus sa stal aj základom pre frekvenčný slovník chorvátčiny, ktorého tvorba sa finalizovala začiatkom deväťdesiatych rokov, [5] avšak kvôli rôznym okolnostiam (aj politickým) tento slovník vyšiel až v r. 1997.

Na dobrú tradíciu korpusovej lingvistiky v Chorvátsku nadväzuje od začiatku deväťdesiatych rokov aj budovanie Chorvátskeho národného korpusu (HNK). HNK sa vypracúva v Jazykovednom ústave Filozofickej fakulty Univerzity v Záhrebe (Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu). Celkový projekt je podporovaný Ministerstvom vedy ako strategický a štátny záujem Chorvátskej republiky. Vzormi pre HNK sú český a britské korpusy [6] a tvoria ho:

- 30M: reprezentatívny 30-miliónový korpus súčasnej chorvátčiny (texty, vzniknuté od roku 1990)
- HETA: Chorvátsky elektronický textový archív tvorený nevyváženou zbierkou korpusov, ktoré sú alebo staršie od r. 1990 alebo nezodpovedajú požiadavkám reprezentatívnosti 30M, ale samy osebe sú niekoľkomiliónovými významnými textovými databázami. Korpusy HETA sú spracované rovnakou metodológiou ako 30M.

Chorvátsky korpus sa zostavoval na základe odporúčaní iniciatívy EAGLES[5], ktoré sa týkajú žánrovej a textovej typológie. Chorváti plánujú jeho široké uplatnenie v tradične gramatických oblastiach (pravopisná problematika, výskum flexie a derivácie), najmä ale v lexikológii a lexikografii (chorvátske výkladové a inojazyčné slovníky, tezaurusy, terminologické a pravopisné slovníky, atď.) a informatike (indexovanie a prehľadávanie textových databáz, výroba počítačových nástrojov na spracovanie prirodzeného jazyka)[6] .

2.9 Korpus srbského jazyka

Projekt korpusu srbského jazyka (CSL) siaha až do roku 1957 keď sa s jeho budovaním začalo na Ústave experimentálnej fonetiky a rečovej patológie v Belehrade v rámci širšieho projektu automatického rozpoznávania textu a strojového prekladu. Početný tím (80 lingvistov a viac ako 300 technických pracovníkov) vedený prof. Đorđe Kostićom vypracoval do r. 1962 originálny gramatický anotačný systém, obsahujúci okolo 2000 značiek a ručne označkoval celý korpus. Projekt sa však po r. 1962 na celých 30 rokov zastavil a k jeho obnoveniu došlo až v r. 1996 vďaka spojenému úsiliu Ústavu experimentálnej

[4] http://www.fida.net/

[5] EAGLES (Expert Advisory Group on Language Engineering Standards) je ďalšia iniciatíva v rámci Európskej únie, ktorá vypracovala odporúčania na kompiláciu, reprezentáciu a lingvistickú anotáciu jazykových dát v elektronickej podobe.

[6] http://www.hnk.ffzg.hr

fonetiky a rečovej patológie a Laboratória experimentálnej psychológie Belehradskej univerzity. Po konverzii textov do elektronického formátu je ďalšou fázou jeho budovania automatické označkovanie pomocou aktualizovaného pôvodného anotačného systému.

Samotný korpus má menší rozsah – 11 miliónov slovných tvarov a skladá sa z piatich podkorpusov zachytávajúcich srbský jazyk od 12. storočia až po súčasnosť. Z toho štyri diachronické korpusy (texty z 12. – 17 storočia, texty z 18. – a prvej pol. 19. storočia, celé dielo Vuka Karadžića, texty z druhej pol. 19. storočia) počítajú 4 milióny slovných tvarov a jeden synchrónny (publicistika, poézia, prózy, esejistika, vedecko-populárna literatúra) má objem 7 miliónov slovných tvarov. Informácie o korpuse sú zverejnené na internete[7], ale zatiaľ nie je dostupná prehľadávateľná časť.

Referencie

[1] Acta Universtitatis Carolinae. Philologica 3-4. *Studie z korpusovej lingvistiky,* 1997.

[2] Piotr Bański. Anotacja zewnętrzna: wpływ architektury korpusu IPI PAN na efektywność jego tworzenia oraz wykorzystania. Złożone do *Poloników,* 2002.

[3] Korczakowska, Monika. Zagadnienia polskich korpusów tekstów. Rukopisný materiál, 2002.

[4] Przepiórkowski, Piotr Bański, Łukasz Dębowski, Elżbietta Hajnicz, Marcin Woliński. Konstrukcja korpusu IPI PAN. Złożone do *Poloników*, 2002.

[5] Marko Tadić. Od korpusa do čestotnog riječnika hrvatskoga književnog jezika. *Rad zavoda za slavensku filologiju* 27:169-178, 1992.

[6] Marko Tadić. Računalna obradba hrvatskih korpusa: povijest stanje i perspektive. *Suvremena lingvistika* 43-44: 388-394, 1997.

[7] Marcin Woliński, Adam Przepiórkowski. Projekt anotacji morfosyntaktycznej korpusu języka polskiego. *IPI PAN Reports 938*, 2001.

[7] http://www.serbian-corpus.edu.yu/