Proměna Českého akademického korpusu *

BARBORA HLADKÁ – JAN KRÁLÍK

The transformation of the Czech Academic Corpus

ABSTRACT: The Czech Academic Corpus was created during the 1970s and 1980s at the Czech Language Institute under the supervision of Marie Těšitelová. The main motivation to build it (a total of 540 thousand word tokens) was to obtain the quantitative characteristics of contemporary Czech. The corpus is structurally annotated on two levels – the morphological level and the syntactical-analytical level. The original stochastic experiments in morphological tagging of Czech were performed using the corpus at the beginning of the 1990s. Given this, the corpus-based processing of Czech was launched. At the end of 1990s, work on the Prague Dependency Treebank had started (independently from the corpus) and its first edition was published in 2001. In considering future released versions of the treebank, we have decided to convert the corpus into the treebank-like format. This article focuses on the twenty-year history of the Czech Academic Corpus. Special attention is devoted to thus far unpublished facts about the corpus annotation. The conversion steps resulting in the first version of the Czech Academic Corpus are described in detail.

Key words: annotated corpus, annotation scheme conversion, natural language processing

Klíčová slova: anotovaný korpus, konverze anotačního schématu, zpracování přirozeného jazyka

Čas a šťastná náhoda si spolu velmi dobře rozumějí, většinou. Uvedou nás do situací, do kterých bychom se dostali za předpokladu opravdu „vychytralé“ intuice, většinou. Představíme jednu takovou situaci, která nese reálnou podobu a reálné jméno – Český akademický korpus. Zmínili jsme čas – proto jej představíme na pozadí této veličiny.

Vrátíme se o dvacet let zpátky do doby, kdy vznikl. Připomeneme dobu před deseti lety, kdy otevřel nové možnosti aplikace stochastických metod v počítačovém zpracování češtiny. Popíšeme jeho současnou příbuznost s Pražským závislostním korpusem.

1. Český akademický korpus před lety

1.1. *Motivace*

Myšlenka Českého akademického korpusu (ČAK), realizovaná v letech 1971–1985 v oddělení matematické lingvistiky v Ústavu pro jazyk český ČSAV, neměla na počátku pevný obrys. Byly tu jednak starší kvantitativní studie ze školy B. Trnky a průlomová technika mechanografické laboratoře J. Štindlové, jednak ambice oddělení, založeného v roce 1961, které po emigraci L. Doležela do Kanady převzala M. Těšitelová. Teprve její zkušenost z ruční práce na rozsáhlém svazku Frekvence slov, slovních druhů a tvarů v českém jazyce (FSSDTČJ, Jelínek – Bečka – Těšitelová, 1961) přinesla myšlenku soustředit nový rozsáhlý textový materiál, netřídit jej však ručně, ale kvantitativní data o frekvencích gramatických a syntaktických kategorií získat pomocí techni-

* Konverze Českého akademického korpusu probíhá v rámci projektu Data a nástroje pro informační systémy, id. č. GA AV 1ET101120413, viz <http://ufal.mff.cuni.cz/rest>.

ky (Těšitelová, 1983a, 1984a; Těšitelová – Uhlířová – Králík, 1984). Výhled směřoval ke kvantifikaci významů – k sémantickému frekvenčnímu slovníku (Těšitelová, 1986; Confortiová, 1990; Ludvíková, 1990; Těšitelová, 1990). První představa počítala s děrnými páskami a s textovým rozsahem jeden milion slov, což byl rozsah materiálu pro frekvenční slovník slovenštiny (Mistrík, 1969) (FSSDTČJ byl založen na textech o celkovém rozsahu 1 623 527 slov). Třídění měly obstarat samočinné počítače, jak se tehdy říkalo. Proti jistotě mechanických třídiček děrných štítků to znamenalo sázku na nekontrolovatelnost funkcí elektroniky, tedy risk. Překvapivé je, že již v tomto stadiu projektu na prahu 70. let se užívalo termínu *korpus*. Neměl být ovšem cílem, ale pouze prostředkem, pracovní fází.

1.2. *Základní charakteristika*

Rozběh projektu ČAK zpomalovaly souběžné rozsáhlé diskuse o pojetí akademické mluvnice češtiny. Padlo proto rychlé rozhodnutí ukončit sérii seminářů k jednotlivým tématům projektu (k vymezení jednotky výběru a textové základny, k metodice přípravy a zpracování dat atd.), dál v původním přístupu mnoho nemodernizovat a zůstat u tradičního, systematicky dobře propracovaného pojetí morfologie a závislostní syntaxe (Šmilauer, 1972). Tím se korpus jako původně zamýšlený doplněk mluvnice vyčlenil z vývojové vlny, osamostatnil se a ve výsledku mohl dokonce mluvnici o několik let předběhnout, ačkoli měl původně být jejím apendixem. Zatímco Mluvnice češtiny 1–3 (Petr, 1986–1987) dospěla k valenčnímu přístupu a k valenčnímu vidění větné syntaxe, pevné zakotvení korpusu v tradičním pojetí gramatiky umožnilo využít relativně ostrých definic gramatických kategorií a ovšem i jednoduššího pohledu na lexikon. Zásady pro zachycení morfologických kategorií vypracovaly M. Těšitelová a M. Ludvíková, zásady pro zachycení syntaktických kategorií L. Uhlířová a I. Nebeská. Návrhem výběru textů byl pověřen J. Kraus a technickou stránkou řešení J. Králík. Myšlenka byla: klasickým rozborem textu identifikovat („ručně“) maximum morfologických a syntaktických informací a ty pak spolu s textem převést „do paměti počítače“ (do elektronické podoby) tak, aby podle nich bylo možno vybírat, sčítat, třídit a uvádět příklady. Cílovou představou byla řada frekvenčních seznamů dat, frekvenčních slovníků a koincidenčních tabulek, v souhrnném pojmenování všestranná kvantitativní analýza současné psané a mluvené češtiny.

1.3. *Vnitřní formát*

Nikoli děrné pásky, ale děrné štítky s osmdesáti sloupci nabídly pro záznam a přenos dat poměrně bohaté možnosti. Aby zůstalo zachováno 26 polí pro tvar slova a 24 polí pro lemma (delší bylo už jen slovo *dodavatelsko-odběratelský*), prostor pro morfologické a syntaktické informace se musel uskrovnit, ale žádný požadavek nakonec nezůstal stranou. Osm míst stačilo pro numerický kód zachycující gramatické kategorie všech ohebných i neohebných druhů slov a na 6 + 8 místech byl ve dvou shlucích zachycen popis charakteristik relevantních pro syntax tak, aby bylo možno rekonstruovat celý graf věty včetně závislostí, vzdáleností, větvení, koordinací i stavby souvětí (viz níže).

U každého slova zůstala jednoznačná numerická lokalizace. Kód (systém značek, numerických „tagů“), který byl rovněž dílem autorského týmu, umožnil zachovat přehlednost a logickou zapamatovatelnost. Je možná paradoxní, že ve stadiu projektu kód nebyl publikován. Předem publikován – vlastně – nebyl ani projekt sám.

1.4. *Výběr textů, styly*

Původní představa uvažovala o podobném výběru textů z oblasti beletrie, publicistiky, naučné literatury, jako tomu bylo u FSSDTČJ, a nově z administrativy. Návrh proporcí zastoupení byl výsledkem šetření, která provedl J. Kraus. Uvnitř stylových oblastí nemělo být užíváno celých děl, jako ve FSSDTČJ, ale výběrů o menším rozsahu. Statistické sondy J. Králíka ukázaly, že pro zachycení struktury slovních druhů a hlavních gramatických kategorií by stačil rozsah 2000 slov. Jenže pojem *Dva tisíce slov* byl v roce 1971 tabu, takže za textovou jednotku byl určen rozsah 3000 slov. V počáteční etapě byly zpracovány texty publicistické. Symbolicky první text pocházel z Rudého práva a rovněž další tituly novin a časopisů byly vybírány podle dobového pohledu (Svět práce, Obrana lidu, Zemědělské noviny, Tribuna, Úder, Svoboda, Pochodeň, Pravda, Průboj). Po jejich vyčerpání byla kritériem výběru dostupnost. Podobně se pak postupovalo u naučných textů (Nadhodnota a její formy, Určování efektivnosti za socialismu, Společenské vědy ve škole, Sociální jistoty včera a dnes, Společenská struktura a revoluce, Ke kritice buržoasních teorií společnosti, Vědeckotechnická revoluce a socialismus) a s nemalou obtíží u textů administrativních (Hospodaření s domovním bytovým majetkem, Pracovní řád, Národní pojištění, Kolektivní smlouvy). Ačkoli byl výběr deformován dobou, dostalo se i na tituly a obory dobou neovlivněné (Jak rozumíme chemickým vzorcům a rovnicím, Škoda 1000 MB, Pražský vodovod, Archeologické nálezy, Nauka o materiálu, Tranzistory řízené elektrickým polem, Elektrotechnický obzor, Stažlivost myokardu, Výzkum hlubinné geologické stavby Československa, Nukleární medicína, Pokroky matematiky, fyziky a astronomie, Hvězdářská ročenka). Největší obtíž nastala při získávání textů mluvených, jichž měla být podle původních představ plná čtvrtina. Zvukové záznamy odborných přednášek a zvláště administrativních povelů a příkazů se získávaly velmi obtížně. Úhrn práce byl časově tak náročný, že z původního projektu bylo nutno rezignovat na beletrii. Z představy o celku jednoho milionu slov tak zůstalo u 540 000 slov tzv. věcného stylu, tj. 60 textů z publicistiky, 100 textů z naučné literatury a 20 textů z administrativy (každý po 3000 slov). Úplný seznam textů je uveden v příloze.

Další postup byl jednotný: vybraný text byl po přepsání na psacím stroji (ob dvě řádky s většími mezerami) opatřen ručně kódem (morfologie modře, syntax červeně), to vše prošlo dvojí revizí a opravená data byla v děrovnách děrovačkami vyděrována na děrné štítky. Z děrných štítků byla data sejmuta mechanickou čtečkou a v elektronické podobě uložena na magnetické pásky. Odtud se provedl kontrolní tisk, který obvykle ukázal nové překlepy (předěry?). Opravy se prováděly zvláštním programem pomocí dalších děrných štítků při kopírování z pásky na pásku, kdy byl chybný záznam vyhledán a nahrazen správným. Teprve pak byl text zařazen do pracovního archivu a zálohován.

1.5. *Technické zázemí*

Protože (sálové) počítače IBM 370 a Tesla 200 v době zahájení projektu nepracovaly běžně s českou abecedou, tisk probíhal ve dvou řádcích, nahoře se v místech písmen s háčky tisklo V, v místech písmen s čárkami lomítko /, a dole text bez diakritiky. Z důvodů fonologicko-„perspektivních“ bylo „ů“ sjednoceno s „ú“ a v zájmu retrográdního třídění byla spřežka „ch“ nahrazena samostatným grafémem (už tehdy se jmenoval zavináč). Pro běžné abecední třídění bylo třeba osmdesátisloupcový záznam rozšířit o dvě další zóny se zvláštními konverzemi tvaru slova a lemmatu na 130 míst. Třídění se provádělo až podle těchto dodatečných zón. Retrográdní třídění vyžadovalo zvláštní úpravu a bylo ještě složitější: trvalo šest hodin a zdařilo se až na potřetí.

Při práci s korigovanými daty se ukázaly jako velmi praktické výpisy kontextů sledovaných jevů (výskytů vybraných lemmat v konkrétních spojeních, výpisy výskytů předložkových vazeb, druhů rozvití, typů věty nebo typů celého souvětí), dále spojení takto získaného materiálu ze všech knihovních pásek v nový pracovní celek (ten se také nazýval korpus), a další třídění, přeuspořádání podle korelací jevů, statistické součty jevů prostých, jejich koincidencí a podmínění apod. Dnes to vše umožňuje např. CD u nového Frekvenčního slovníku češtiny (Čermák – Křen, 2004). Tehdy bylo nutno každý takový požadavek řešit novým počítačovým programem. Každá maličkost se programovala v jazycích APS a ASSEMBLER, ladění probíhalo jen pomocí výpisů a nových štítků, v lepším případě komunikací s počítačem pomocí elektrického psacího stroje. Výsledné tisky abecedních a frekvenčních seznamů, spekter tvarů (s frekvencemi), systematické soupisy podle „kódů“ (= tagů = gramatických a syntaktických značek), atd. atd., představovaly téměř nepřeberný materiál pro kvantitativní popis fungování češtiny z velmi rozmanitých hledisek.

1.6. *Porovnání s jinými korpusy té doby*

V době vzniku ČAK se neuvažovalo o tom, že by data sama mohla mít svou nezávislou hodnotu i ve vzdálenější budoucnosti. Uchování na velkých magnetických páskách bylo tak jako tak neoperativní a kopie se pořizovaly jen pro zálohování. Ačkoli ČAK byl jediným tak podrobně značkovaným korpusem flektivního jazyka, zdálo se, že důležitější než data sama bude jejich bezprostřední exploatace.

V první vlně využití ČAK bylo v letech 1980–1986 publikováno šest frekvenčních slovníků (Těšitelová, 1980a, 1980b, 1982c, 1983c; Těšitelová – Petr – Králík, 1985; Těšitelová – Petr – Králík, 1986a), tři samostatné svazky tabulek a grafů (Těšitelová, 1982b, 1983d, 1984b), další svazky seznamů, soupisů a dat o publicistice (Těšitelová, 1981, 1982a), naučných textech (Těšitelová, 1983b, 1983f) a administrativním stylu (Těšitelová, 1985b). Došlo i na celou řadu dílčích článků zejména o morfologii (Ludvíková, 1983, 1986) a syntaxi (Uhlířová – Nebeská – Králík, 1982; Nebeská, 1983, 1986; Uhlířová, 1983, 1986, 1990) a o dalších tématech (Těšitelová, 1979; Confortiová, 1983, 1986; Králík, 1983a, 1983b; Těšitelová, 1983e, 1985a, 1987, 1992; Těšitelová – Petr – Králík, 1986b; Nebeská, 1986, 1990; Králík, 1991). Hlavní souhrn dat byl

zveřejněn v monografii Kvantitativní charakteristiky současné češtiny nakladatelstvím Academia (Těšitelová, 1985c). Popis technických předpokladů, zásad a příkladů řešení zabral další samostatný svazek (Králík, 1987). V relaci k ostatním byly v této době nejméně využity možnosti uplatnění dat o větě a souvětí. Kód např. umožňoval zápis závislostních grafů v lineární podobě a tím rychlé třídění a součty. Publikování se nedočkaly.

V době svého vzniku byl ČAK jediným světovým korpusem anotovaným na morfologické i syntakticko-analytické rovině. O to smutněji vyznívá fakt, že se mu nedostalo takové světové popularity, jakou by si zasloužil. Pro úplnost dodáváme, že „konkurenty" byly korpusy americké a britské angličtiny – Brownův korpus (Francis – Kucera, 1979) a LOB korpus (Atwell – Leech – Garside, 1984), které obsahovaly morfologicky anotovaný jeden milion slov textů z roku 1961.

1.7. *Archiv zdrojů*

Rukopisné zdroje ČAK zůstaly po nějaký čas v archivu, i když bylo oddělení matematické lingvistiky v roce 1985 zrušeno. Zájem o archiv ovšem opadl. Nikoli zájem o elektronickou podobu dat. Ta byla již v pracovním stadiu využita k testování vyhledávacích programů např. v ČTK. Rychlý technický pokrok, změny formátů záznamu i archivačních nosičů a přechod od sálových počítačů k PC byly kolem roku 1989 tak překotné, že se jen s vypětím dařilo data ČAK stejně rychle konvertovat a ukládat do všech přicházejících nových podob. Nebylo totiž jisté, která z nich bude perspektivní, a která ne. Záchrana se nakonec zdařila a dnes jsou data konvertována do kompatibilního formátu, zálohována již na několika místech i v zahraničí a snadno dostupná na CD.

2. Premiéra korpusových metod v počítačovém zpracování češtiny před deseti lety

Počátek devadesátých let minulého století se ukázal jako zásadní z pohledu počítačového zpracování češtiny s využitím korpusů. V té době totiž J. Hajič pobýval na stáži ve výzkumném centru IBM ve Spojených státech amerických (IBM T. J. Watson Research Center, NY). Pracoval ve skupině vedené F. Jelinkem, která řešila úlohu statistického překladu z angličtiny do francouzštiny. Stochastický (statistický) pohled na věc se dá ilustrovat následovně: předpokládejme, že nevíme, jaká anglická věta se překládá, nicméně máme před sebou její francouzský překlad; k tomuto překladu hledáme anglickou větu, z které nejpravděpodobněji vznikl – v ideálním případě nalezneme tu, kterou měl mluvčí na mysli. Jednou z cest, jak hledat nejpravděpodobnější anglické věty, je nejdříve přečíst několik (čím více, tím lépe) anglicko-francouzských paralelních textů a zapamatovat si možné francouzské (anglické) překlady anglických (francouzských) vět, či úseků vět, včetně četností, s jakou jednotlivé překlady nastaly. Následně se prohledávají zapamatované francouzské věty včetně jejich anglických překladů. Francouzská věta, jejíž anglický překlad se nejvíce blíží vstupní anglické větě, je vybrána.

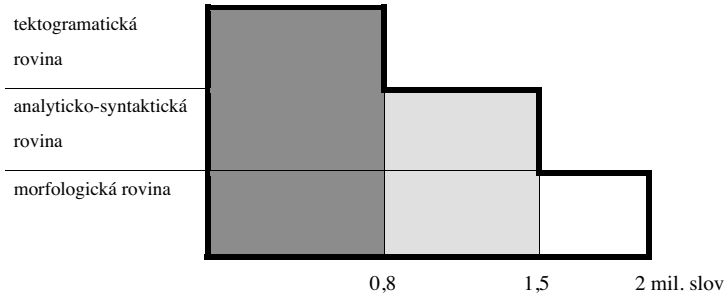
Po návratu J. Hajiče, zaujatého statistickými metodami v počítačové lingvistice, bylo možné jeho zájem ještě posílit. Poprvé jsme mohli realizovat myšlenku statistického přístupu v rámci počítačového zpracování češtiny – konkrétně v úloze morfologického

značkování, tedy automatického určování slovních druhů a tvaroslovných kategorií pro slova textu (Hladká, 1994; Hajič – Hladká, 1997a; Hajič – Hladká, 1997b; Hladká, 2000). Jako trénovací data v experimentech byla použita právě data z ČAK. Tím se čeština připojila k ostatním, z geografického pohledu západním jazykům, které ve svém počítačovém zpracování byly v té době mnohem dál.

Postupně se pracovalo na vylepšení automatického morfologického značkování, nicméně na svět se začala, nezávisle na ČAK, prodírat myšlenka vybudování anotovaného korpusu češtiny. ČAK byl ponechán stranou a v roce 1996 se začal anotovat Pražský závislostní korpus (PZK, viz <http://ufal.mff.cuni.cz/pdt>).

3. Český akademický korpus dnes

V současné době se PZK pyšní již dvěma verzemi. První verze byla publikována v roce 2001 (Hajič et al., 2001) s celkovým objemem dva miliony anotovaných slov – kompletně morfologicky (2 mil. slov, Hana – Zeman, 2005) a částečně syntakticko-analyticky (1,5 mil. slov). Za čtyři roky přibyly ještě u části dat anotace na rovině významové, tzv. tektogramatické (0,8 mil. slov). Druhá verze bude oficiálně publikována v roce 2006. Rozložení dat (viz obr. 1) anotovaných na jednotlivých rovinách odpovídá směru, kterým anotování probíhalo – od jednoduššího ke složitějšímu. Tedy od roviny morfologické k rovině tektogramatické přes rovinu syntakticko-analytickou.



Obr. 1: Rozložení objemu dat PZK 2.0 dle rovin anotování

Téměř desetileté anotační úsilí s sebou přineslo zkušenosti, které se pojí jenom se superlativy. A vyzbrojeni právě touto zkušeností jsme si vzpomněli na ČAK a rozhodli jsme se ho vzhledem k jeho nezanedbatelnému objemu (cca 550 tis. slov) převést do podoby kompatibilní s PZK. Kompatibilitou (onou příbuzností zmíněnou v samotném úvodu) myslíme shodu ve vnitřním formátu a v anotačních schématech. Tabulka 1 soustřeďuje základní porovnání anotací a textů v PZK 2.0 a ČAK.

V následující části této statě uvedeme vše, co v sobě kompatibilita s PZK ukrývá. Aby odhalování bylo srozumitelnější, formulujeme pět základních fází, kterými prošlo anotování PZK. Fáze po sobě následovaly tak, jak jsou uvedeny, až na fáze c) a d), které nemohly být realizovány jinak než paralelně. I při kontrolách anotací ve fázi e) se ještě dolaďovaly pokyny:

a) promyšlení strategie anotování,
b) výběr týmu anotátorů,
c) formulace pokynů pro anotátory,
d) anotování,
e) kontrola anotací.

V dané situaci, kdy pracujeme s již anotovaným korpusem, můžeme charakterizovat fáze anotačního procesu následovně:

f) promyšlení strategie konverze vnitřních formátů a anotačních schémat,
g) výběr týmu anotátorů,
h) anotování,
i) kontrola anotací.

Strategie anotování a) je nahrazena strategií konverzní procedury f) se známými vstupními parametry – anotační schéma ČAK a PZK a vnitřní formát ČAK a PZK. Pokud by se ukázalo, že danou konverzi je možné realizovat zcela automaticky, tedy že jak vnitřní formáty, tak i anotační schémata si jednoznačně odpovídají „jedna k jedné“, fáze výběru anotátorů by nebyla nutná a fáze vlastního anotování h) by se provedla automaticky. Tato konstelace bohužel nenastala. Formulace pokynů pro anotátory nebyla nutná, protože již existovaly pokyny pro anotování PZK. Kontrola anotací v rámci zajištění co nejvyšší konzistence je naopak nedílnou, velmi časově náročnou součástí jakéhokoli anotačního procesu. Dílčí kroky konverze, které vyústí publikováním ČAK verze 1.0, byly realizovány v tomto pořadí:

1. Konverze vnitřního formátu – viz část 3.1.,
2. Konverze morfologických anotací – viz část 3.2.,
3. Korektury – viz část 3.3.,
4. Automatické kontroly morfologických anotací – viz část 3.4.

Konverze syntakticko-analytických anotací a větných anotací bude probíhat až po vydání ČAK 1.0. Koncepčně bude zcela jistě odpovídat krokům f)–i).

Tab. 1: Základní porovnání materiálu v PZK 2.0 a ČAK

CHARAKTERISTIKA	PZK 2.0		ČAK	
	POČET SLOV	POČET VĚT	POČET SLOV	POČET VĚT
ANOTACE MORFOLOGICKÁ	2 mil.	116 tis.	550 tis.	31 tis.
ANOTACE SYNTAKTICKO-ANALYTICKÁ	1,5 mil	88 tis.	550 tis.	31 tis.
ANOTACE VĚTNÁ	—	—	550 tis.	31 tis.
ANOTACE TEKTOGRAMATICKÁ	0,8 mil.	49 tis.	—	—
	DOKUMENTY		DOKUMENTY	
TEXTY NOVINOVÉ	81 %		33 %	
TEXTY ADMINISTRATIVNÍ	—		11 %	
TEXTY ODBORNÉ	9 %		56 %	
	DOKUMENTY		DOKUMENTY	
TEXTY PÍSEMNÉ	100 %		75 %	
TEXTY MLUVENÉ	—		25 %	

3.1. *Konverze vnitřního formátu*

Původní vnitřní formát ČAK (viz též část 1.3.) je přirozeně jednoduchý, tzv. „sloupcový“ – každému slovu textu odpovídá jeden řádek, na kterém kromě daného slova jsou i ručně přiřazené morfologické anotace, lemma, analytické anotace a větné anotace spolu s pořadím slova ne v rámci kompletního ČAK, ale v rámci daného souboru. Sloupce s morfologickou značkou, syntakticko-analytickou značkou, větnou značkou a s pořadím slova jsou fixní délky. Vše je názorně vidět na příkladu věty *Sto lidí, sto povah, názorů, sklonů a zájmů*:

Tab. 2: Věta *Sto lidí, sto povah, názorů, sklonů a zájmů* v původním formátu ČAK

SLOVNÍ FORMA	MORFOLOGICKÁ ZNAČKA[1] (8 pozic)	LEMMA	SYNTAKTICKO-ANALYTICKÁ ZNAČKA[2] (9 pozic)	VĚTNÁ ZNAČKA[3] (6 pozic)	POŘADÍ (7 pozic)
sto	140411	sto	61	911	0250090
lidí	110122	lid	31_01		0250091
sto	140411	sto	61_021		0250092
povah	110322	povaha	31_01		0250093
názorů	110222	názor	31_011		0250094
sklonů	110222	sklon	31_011		0250095
a	81	a			0250096
zájmů	110222	zájem	31_021		0250097

Volba takto jednoduchého formátu odpovídá době vzniku korpusu a stupni vývoje informatiky samotné. PZK, který vznikl o dvacet let později, je co do vnitřního formátu podstatně bohatší – je spojen se třemi formáty – CSTS[4], fs[5] a PML[6]. Formát fs je originálně pražský, formáty CSTS a PML jsou také „pražské“, ale vycházejí ze světových standardů – CSTS z SGML, PML z XML.

Pozornému čtenáři jistě neuniklo, že v reprezentaci věty v tab. 2 chybí interpunkce. Nejedná se o opomenutí autorů článku, ale o skutečnou vlastnost korpusu. Připomeňme, že hlavní motivací pro anotování ČAK bylo získat frekvenční charakteristiky češtiny té doby ze skutečného materiálu (viz část 1.1.). Z tohoto pohledu nebyla ani interpunkce, ale ani ciferné výrazy zajímavé. Proto byly z původních textů odstraněny – viz např. věta *V tomto směru počítá koncepce severomoravského kraje do roku [cifra] s realizací [cifra] akcí.*

V době konverze byl formát PML ve fázi testování, proto jsme pro konverzi zvolili formát CSTS. Při tomto kroku jsme se potýkali s problémy spíše technického charakteru, které vyžadovaly ruční zásah. Jakékoli změny v datech byly dokumentovány v CSTS elementech, aby se vždy dala rekonstruovat původní podoba korpusu. Hranice

[1] Viz popis na adrese <http://ufal.ms.mff.cuni.cz/REST/CAC/tOrig.html>.
[2] Viz popis na adrese <http://ufal.ms.mff.cuni.cz/REST/CAC/aOrig.html>.
[3] Viz popis na adrese <http://ufal.ms.mff.cuni.cz/REST/CAC/sOrig.html>.
[4] Viz popis na adrese <http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/csts/html/DTD-HOME.html>.
[5] Viz popis na adrese <http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/fs/index.html>.
[6] Viz popis na adrese <http://ufal.ms.mff.cuni.cz/pdt2.0/doc/data-formats/pml/index.html>.

vět bylo možné jednoznačně rekonstruovat na základě větných anotací. Ovšem na hranice odstavců a dokumentů jsme museli rezignovat, protože tato informace nebyla v datech uchována.

3.2. *Konverze morfologických anotací*

Morfologické anotace se skládají z lemmatu a morfologické značky (Hajič, 2004). Pokud si vstupní a výstupní anotační schémata neodpovídají jedna k jedné, bez systému automatické morfologické analýzy by kontrola převedených morfologických anotací nebyla možná.

Konverze morfologických anotací byla zahájena převedením původních morfologických značek do anotačního schématu PZK. Původní lemma bylo ponecháno. Následovala automatická morfologická analýza, která přiřadila každému slovu textu sadu možných lemmat s příslušnými morfologickými značkami dle konvence zavedené v PZK.

Slova (slovní formy), která morfologická analýza nezná (tj. neumí je analyzovat), musela být kontrolována ručně. Pro ostatní slova byly původní ruční anotace ČAK automaticky, velice opatrně porovnávány s výstupem morfologické analýzy. V případě jednoznačné shody ruční anotace s právě jednou dvojicí (lemma, značka), kterou nabídla analýza, byla ruční anotace klasifikována jako správná a dalším kontrolám již nepodléhala. Podobně tomu bylo i v případě, kdy si morfologické značky odpovídaly a ruční lemma se od lemmatu vygenerovaného morfologickou analýzou lišilo pouze v jeho syntakticko-sémantických příznacích. Pokud tyto jednoznačné situace nenastaly, bylo dané slovo zpracováno ručně s tím, že anotátor měl k dispozici jak původní ruční anotaci, tak i výstup analýzy. Dle kontextu vybral buď jednu z možností nabízenou analýzou, nebo opravil původní ruční anotaci, pokud ani jedna z nabízených možností analýzy nebyla vzhledem ke kontextu správná. Nejčastěji se vyskytly následující kolize:

– velká, malá písmena. V ČAK byla všechna slova uvedena s malým počátečním písmenem. Z původních morfologických značek se dala jednoznačně detekovat vlastní jména, ale pouze substantiva. Mnohdy bylo možné změnu počátečního písmene vyčíst z porovnávací procedury.
– *ú*/*ů*. V ČAK bylo výhradně používáno *ú.* Všechny jeho výskyty uprostřed slov byly nahrazeny *ů*. Případy, kdy *ú* nemělo být nahrazeno *ů*, byly řešeny opět v rámci porovnávací procedury.
– některá slova nebyla v ČAK anotována, pouze měla přiřazené pořadové číslo. V PZK je zvolena taková strategie, že každému slovu je přiřazena morfologická anotace. Proto byly doplněny anotace i k původně neanotovaným slovům. Pro ilustraci uvádíme několik vět – podtržená slova jsou právě ta, která nebyla v ČAK anotována: *Chystají se dokumentární filmy. Tuto funkci by měla plnit každoroční, celoslovenská konference kulturních pracovníků profesionálů a amatérů. Federální ministerstvo vnitra stanoví v dohodě se zúčastněnými ústředními orgány…* (původní lemma u *v* je *v dohodě s* a slova *dohodě* a *s* neměla přiřazená žádné lemma)

3.3. *Korektury*

Během doposud prováděných akcí jsme nepřítomnost interpunkce a ciferných výrazů dokázali obejít. Avšak s ohledem na konverze stromových struktur, na plán přičlenit ČAK k PZK a použít ho jako další materiál pro trénování algoritmů strojového učení se jedná o překážky natolik zásadní, že bylo nutné je vyřešit. Protože původní textové zdroje jsou již nedostupné, nabízelo se jediné řešení, a to všechny dokumenty ČAK přečíst a provést patřičné korektury: kontrola velkých/malých počátečních písmen, označení nesrozumitelného textu, označení místa, kde chybí ciferný výraz, určení typu chybějícího ciferného výrazu, označení místa, kde chybí interpunkční znaménko.

Na korekturách pracovali studenti filologických oborů s podporou uživatelsky příjemného nástroje. Skupina korektorů byla osmičlenná a každý soubor ČAK byl zpracován dvěma korektory. Následně bylo vyhodnoceno, jak se korektoři ve dvojicích ve svých rozhodnutích shodovali, příp. neshodovali. Míra shody byla relativně nízká, proto musely být vyřešeny diskrepance mezi korektory opět ručně.

Od okamžiku, kdy jsme se rozhodli chybějící informace do textů doplnit, jsme si plně uvědomovali, že se do textů vkládají názory korektorů. V takovém případě je potřeba myslet především na konzistenci vložených názorů. Tu jsme se pokusili zajistit právě tím, že na závěrečném vyřešení diskrepancí pracovali pouze dva korektoři, a to ti, kteří měli v prvním kole nejvyšší míru shody. Nikdo třetí do jejich rozhodování již nezasahoval. Korektury jsme po druhém kole ukončili, protože ambice na stoprocentní mezikorektorskou shodu neměly své opodstatnění. Jediným cílem korektur bylo doplnit chybějící informace tak, abychom získali syntakticky správné věty.

3.4. *Kontroly morfologických anotací*

Při zpracování textu automatickou morfologickou analýzou může nastat situace, že analýza danou slovní formu nezná, nebo ani jedna z možností analýzy není správná vzhledem ke kontextu. V obou případech je nutná ruční anotace, při které přirozeně nastanou chyby, ať už na úrovni lemmat, či značek. Automatickými procedurami je možné většinu chyb (nesrovnalostí) detekovat, ovšem jejich odstranění je vázáno opět na ruční zásah, protože těch případů, kdy se jedná o systematickou chybu v anotování, je velmi malé procento – jak ukazují zkušenosti s kontrolou PDT 2.0.

V rámci kontrol morfologických anotací PDT 2.0 byla sestavena celá řada kontrolních skriptů, které využívaly i anotace z vyšších rovin (pokud existovaly). Pro kontrolu morfologických anotací ČAK 1.0 byly použity pouze ty skripty, které při hledání kandidátů na chyby využívají informaci dostupnou z roviny morfologické. Pro ilustraci uvádíme několik z nich:

– lemma slovní formy porušuje korektnost lemmat;
– značka slovní formy porušuje korektnost značek, tj. značka musí mít právě 15 znaků a na každé pozici musí být znak z množiny povolených znaků pro tuto pozici; značka je přípustná v morfologickém systému češtiny;

– dvě různé slovní formy mají stejné lemma a značku;
– všechny výskyty jednoho lemmatu nemají značku se stejným slovním druhem;
– lemma slovní formy anotované jako adjektivum nekončí na *ý*, *í*.

4. Závěr

Představili jsme dvacetiletou historii korpusu, která byla započata Korpusem věcného stylu (původní jméno) a která končí první verzí Českého akademického korpusu, z pohledu vnitřního formátu a anotačního schématu pro morfologickou rovinu zcela kompatibilní s projektem Pražského závislostního korpusu. Doufáme, že jsme tím dali dostatečně najevo, jak moc si vážíme úsilí vynaloženého při práci na ČAK v osmdesátých letech 20. století. Snad tím přispějeme i ke světové popularitě ČAK, která ovšem měla přijít o dvacet let dříve.

V druhé polovině roku 2006 bude vydáno CD-ROM s první verzí ČAK. Samotný korpus budou doprovázet i nástroje pro jeho prohlížení a pro vyhledávání těch informací, které může morfologicky anotovaný korpus nabídnout. ČAK bude rovněž přičleněn k datům, která se používají jako data trénovací pro metody strojového učení konkrétně aplikované na úlohu morfologického značkování (určování slovního druhu a hodnot relevantních morfologických kategorií slov textu). Vzhledem k objemu korpusu (550 tis. slov) očekáváme poměrně významné zvýšení úspěšnosti těchto metod.

Publikováním první verze ČAK se jeho historie nepřestává psát. Naopak. Již byla zahájena konverze syntakticko-analytických anotací, která bude publikována opět ve formě CD-ROM jako druhá verze ČAK.

5. Poděkování

Na konverzi Českého akademického korpusu intenzivně spolupracovali kolegové Jan Hajič, Jiří Hana a Emil Jeřábek. Za jejich úsilí jim patří obrovský dík.

LITERATURA

ATWELL, E. – LEECH, G. – GARSIDE, R. (1984): Analysis of the LOB Corpus: Progress and Prospects. In: J. Aarts – W. Meijs (eds.), *Corpus Linguistics*. Amsterdam: Rodopi, s. 40–52.

CONFORTIOVÁ, H. (1983): On prepositions in non-fiction style. In: E. Hajičová – V. Hrabě – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 8*. Praha: Academia, s. 31–42.

CONFORTIOVÁ, H. (1986): On the semantic analysis of prepositions from the quantitative point of view. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 9*. Praha: Academia, s. 53–64.

CONFORTIOVÁ, H. (1990): On the problems of the semantics of Czech adjectives from the quantitative point of view. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 10*. Praha: Academia, s. 25–48.

ČERMÁK, F. – KŘEN, M. (2004): *Frekvenční slovník češtiny*. Praha: Nakladatelství Lidové noviny.

FRANCIS, W. N. – KUCERA, H. (1979): *Manual of Information to Accompany a Standard Sample of Present-Day Edited American English, for Use with Digital Computers*. Providence: Department of Linguistics, Brown University.

HAJIČ, J. (2004): *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Prague: Karolinum.

HAJIČ, J. – HAJIČOVÁ, E. – PAJAS, P. – PANEVOVÁ, J. – SGALL, P. – VIDOVÁ HLADKÁ, B. (2001): *Prague Dependency Treebank 1.0* [CD-ROM]. Linguistic Data Consortium.

HAJIČ, J. – HLADKÁ, B. (1997a): Probabilistic and rule-based tagger of an inflective language: a comparison. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Washington, s. 111–118.

HAJIČ, J. – HLADKÁ, B. (1997b): Morfologické značkování korpusu českých textů stochastickou metodou. *Slovo a slovesnost*, 58, s. 288–304.

HANA, J. – ZEMAN, D. (2005): *Manual for Morphological Annotation, Revision for the Prague Dependency Treebank 2.0 (technická zpráva TR-2005-25)*. Praha: Ústav formální a aplikované lingvistiky MFF UK.

HLADKÁ, B. (1994): *Software Tools for Large Czech Corpora Annotation*. MSc thesis. Prague: Charles University.

HLADKÁ, B. (2000): *Czech Language Tagging*. PhD thesis. Prague: Charles University.

JELÍNEK, J. – BEČKA, J. V. – TĚŠITELOVÁ, M. (1961): *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha: Státní pedagogické nakladatelství.

KRÁLÍK, J. (1983a): Some notes on the frequency – rank relation. In: E. Hajičová – V. Hrabě – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 8*. Praha: Academia, s. 67–80.

KRÁLÍK, J. (1983b): Statistika českých grafémů s využitím moderní výpočetní techniky. *Slovo a slovesnost*, 46, s. 295–304.

KRÁLÍK, J. (1987): *Kapitoly o výpočetní technice: K problémům komunikace lingvista – programátor – počítač*. Praha: Ústav pro jazyk český ČSAV.

KRÁLÍK, J. (1991): Probabilistic scaling of texts. In: R. Köhler – B. R. Rieger (eds.), *Contributions to Quantitative Linguistics*. Dordrecht – Boston – London: Kluwer Academic Publisher, s. 227–240.

LUDVÍKOVÁ, M. (1983): Quantitative aspects of verb categories (based on present-day Czech non-fiction texts). In: E. Hajičová – V. Hrabě – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 8*. Praha: Academia, s. 19–30.

LUDVÍKOVÁ, M. (1986): On the semantics of pronominal adverbs from the quantitative aspect. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 9*. Praha: Academia, s. 43–52.

LUDVÍKOVÁ, M. (1990): Some specific features of the semantics of adverbs. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 10*. Praha: Academia, s. 49–64.

MISTRÍK, J. (1969): *Frekvencia slov v slovenčine*. Bratislava: Slovenská akadémia vied.

NEBESKÁ, I. (1983): Compound/complex sentences in non fiction texts. In: E. Hajičová – V. Hrabě – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 8*. Praha: Academia, s. 53–66.

NEBESKÁ, I. (1986): A contribution to the semantics of modal verbs from the quantitative point of view. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 9*. Praha: Academia, s. 31–42.

NEBESKÁ, I. (1990): On expressing possibility and necessity in Czech. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 10*. Praha: Academia, s. 75–84.

PETR, J. (ed.) (1986–1987): *Mluvnice češtiny, 1–3*. Praha: Academia.

ŠMILAUER, V. (1972): *Nauka o českém jazyku*. Praha: Státní pedagogické nakladatelství.

TĚŠITELOVÁ, M. (1979): On quantitative linguistics in Czechoslovakia. *ITL: Tijdschrift van het Instituut voor Toegepaste Linguēstiek*, 43, s. 53–73.

TĚŠITELOVÁ, M. (1980a): *Frekvenční slovník současné administrativy*. Praha: Ústav pro jazyk český ČSAV.

TĚŠITELOVÁ, M. (1980b): *Frekvenční slovník současné české publicistiky*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1981): On the language of the present-day publicist prose. In: J. Horecký – M. Těšitelová – J. Machek (eds.), *Prague Studies in Mathematical Linguistics, 7*. Praha: Academia, s. 9–26.
TĚŠITELOVÁ, M. (1982a): *Linguistica, II: Kvantitativní charakteristiky současné české publicistiky*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1982b): *Linguistica, III: Kvantitativní charakteristiky současné české publicistiky: tabulky a grafy*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1982c): *Frekvenční slovník současné odborné češtiny*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1983a): On the state of Quantitative Linguistics in studies of Czech. *The Prague Bulletin of Mathematical Linguistics*, 40, s. 15–30.
TĚŠITELOVÁ, M. (1983b): Some quantitative characteristics of non-fiction texts in present-day Czech. In: E. Hajičová – V. Hrabě – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 8*. Praha: Academia, s. 9–18.
TĚŠITELOVÁ, M. (1983c): *Frekvenční slovník češtiny věcného stylu*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1983d): *Kvantitativní charakteristiky gramatických jevů v současné administrativě: tabulky*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1983e): *Linguistica, IV: Psaná a mluvená odborná čeština z kvantitativního hlediska: v rámci věcného stylu*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1983f): *Linguistica, VII: Kvantitativní charakteristiky současné odborné češtiny: v rámci věcného stylu*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1984a): Kvantitativní analýza češtiny s pomocí moderní výpočetní techniky. *Naše řeč*, 67, s. 47–50.
TĚŠITELOVÁ, M. (1984b): *Kvantitativní charakteristiky gramatických jevů v češtině věcného stylu: tabulky a přehledy*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1985a): K využití statistických metod v kombinaci s retrográdním uspořádáním jazykových jednotek. *Slovo a slovesnost*, 46, s. 109–118.
TĚŠITELOVÁ, M. (1985b): *Linguistica, XV: Současná česká administrativa z hlediska kvantitativního*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. (1985c): *Kvantitativní charakteristiky současné češtiny*. Praha: Academia.
TĚŠITELOVÁ, M. (1986): On semantic quantitative analysis. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 9*. Praha: Academia, s. 9–18.
TĚŠITELOVÁ, M. (1987): Kvantitativní lingvistika a počítače. In: M. Těšitelová, *Kvantitativní lingvistika*. Praha: Státní pedagogické nakladatelství, s. 140–143.
TĚŠITELOVÁ, M. (1990): On semantics of nouns from the quantitative point of vew. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 10*. Praha: Academia, s. 9–24.
TĚŠITELOVÁ, M. (1992): *Quantitative Linguistics*. Amsterdam – Philadalphia: John Benjamins.
TĚŠITELOVÁ, M. – PETR, J. – KRÁLÍK, J. (1985): *Retrográdní slovník tvarů adjektiv v současné češtině*. Praha: Ústav pro jazyk český ČSAV.
TĚŠITELOVÁ, M. – PETR, J. – KRÁLÍK, J. (1986a): *Retrográdní slovník současné češtiny*. Praha: Academia.
TĚŠITELOVÁ, M. – PETR, J. – KRÁLÍK, J. (1986b): On some issues of the reverse dictionary of words and forms. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 9*. Praha: Academia, s. 65–74.
TĚŠITELOVÁ, M. – UHLÍŘOVÁ, L. – KRÁLÍK, J. (1984): K automatickému zpracování textu při kvantitativní analýze přirozeného (českého) jazyka. *Slovo a slovesnost*, 45, s. 145–150.
UHLÍŘOVÁ, L. (1983): Simple sentence structure from the quantitative point of view (based on present-day Czech non-fiction texts). In: E. Hajičová – V. Hrabě – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 8*. Praha: Academia, s. 43–52.

UHLÍŘOVÁ, L. (1986): On verbal semantics from the quantitative point of view. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 9*. Praha: Academia, s. 19–30.
UHLÍŘOVÁ, L. (1990): Beginning and the end of sentence: a quantitative study on the present-day Czech. In: E. Hajičová – J. Horecký – M. Těšitelová (eds.), *Prague Studies in Mathematical Linguistics, 10*. Praha: Academia, s. 65–74.
UHLÍŘOVÁ, L. – NEBESKÁ, I. – KRÁLÍK, J. (1982): Computational data analysis for syntax. In: J. Horecký (ed.), *COLING 82*. Amsterdam – New York – Oxford: Elsevier, s. 391–396.

Ústav formální a aplikované lingvistiky MFF UK
Malostranské nám. 25, 118 00 Praha 1
<hladka@ufal.mff.cuni.cz>

Ústav pro jazyk český AV ČR
Letenská 4, 118 51 Praha 1
<kralik@ujc.cas.cz>

Příloha: Soupis textů v Českém akademickém korpusu

Publicistika: písemné texty

Ahoj na sobotu
Čelákovický zpravodaj
Čs. rozhlas I.
Čs. rozhlas II.
Čs. sport
Film a doba
Gramorevue G 73
Haló sobota
Horník a energetik
Chovatel
Kino
Krkonošská Pravda
Květy
Lidová demokracie
Melodie
Městský zpravodaj Brandýsa nad Labem
Mladá fronta
Mladý svět
Naše rodina
Nová Svoboda
Nové Hradecko
Nové Klatovsko
Obrana lidu
Pochodeň
Práce
Pravda
Průboj
Rudé právo
Sázavan
Signál
Služba lidu
Stadión
Stráž lidu
Svět práce
Svět socialismu
Svoboda
Svobodné slovo
Školství a věda
Technický týdeník
Tribuna
Týdeník aktualit
Úder
Večerní Praha
Věda a technika mládeži
Vlasta
Záběr
Zahrádkář
Zahradnické listy
Zápisník Z'73
Zbrojovák
Zemědělské noviny
Zpravodaj TIBY

Publicistika: mluvené texty

Rozhlasové reportáže a rozhovory

Odborný styl: písemné texty

Alpinkářův svět
Arbitrážní praxe
Astronomie
Bezpečnost elektrických spotřebičů
Česká literatura I.
Česká literatura II.
Český lid
Čs. informatika
Čs. psychologie
Dějiny české hudební kultury
Elektronický obzor 6/1974
Elektrotechnický obzor
Filosofický časopis 5/1974
Hospodářské právo
Humanismus v naší filosofické tradici
Hutnictví a strojírenství
Hvězdářská ročenka
Jak na práce s kovem
Jak na práce se stavebninami
Jak rozumíme chemickým vzorcům a rovnicím
K biologickým a psychologickým zřetelům výchovy
Ke kritice buržoasních teorií společnosti
Konflikty mezi lidmi
Lékařská fyzika
Mineralogie
Motivace lidského chování
Nadhodnota a její formy
Národopisné aktuality
Nauka o materiálu
Nukleární medicína
Obkládáme interiéry a fasády
Opravujeme a modernizujeme rodinný domek
Otázky lexikální statistiky
Památková péče 4/1974
Plazma, čtvrté skupenství hmoty
Podstata hypnózy a spánek
Poetika
Pokroky matematiky, fyziky a astronomie
Politická ekonomie
Polovodičová technika
Pražský vodovod
Pro půvab a eleganci
Přirozený jazyk v informačních systémech
Ptáci
Rozvoj osobnosti a slovesné umění
Slovo a slovesnost 4/1973
Sociální jistoty včera a dnes
Sociologický časopis 3/1973
Spisovný jazyk v současné komunikaci
Společenská struktura a revoluce
Společenské vědy ve škole 2/1974
Společnost – vzdělání – jedinec
Stažlivost myokardu
Škoda 1000 MB
Škola opora socialismu
Teorie a empirie
Teorie a počítače v geofyzice
Teplárenství
Tisíciletý vývoj architektury
Tranzistory řízené elektrickým polem
Určování efektivnosti za socialismu
Vědeckotechnická revoluce a socialismus
Vědecko-technický rozvoj za socialismu
Vlastivědný sborník moravský
Výzkum hlubinné geologické stavby Československa
Základní a rekreační tělesná výchova 10/1974
Záruční lhůty potravinářských výrobků
Zesilovače se zpětnou vazbou

Odborný styl: mluvené texty

Archeologické nálezy v Toušeni (Jaroslav Špaček)
Česká filharmonie hraje a hovoří (Václav Neumann)
Divadelní přehlídka
Dlouhodobé skladování masa
Filosofie fyziky (RNDr. Jiří Mrázek, CSc.)
Modelování diod
O počtu koster jednoho grafu
O výchově socialistické inteligence
O vývoji knihovnictví
Obecné otázky jazykové kultury
Ochrany v průmyslových závodech
Opera o Bratrech Karamazových (prof. dr. Václav Holzknecht)
Organizace a řízení vnitřního obchodu
Personalistika
Petrologie sedimentů a reziduálních hornin
Plenární schůze ROH / Pauzy váhání
Práce se čtenářem
Problémy aerodynamiky závodních vozů
Provozní kontrola potrubí
Přednáška o geografii
Přenosové parametry
Působení hromadných sdělovacích prostředků
Rozbor situace v JZD

Seminář o fotografii
Seminář o houbách
Schůze vědecké rady ČSTV
Statické zajištění domu U rytířů
Streptokoky
Úvod do dějin feudalismu
Výklad Zákoníku práce
Základní podmínky pro pěstování zeleniny
Zpráva o cestě do Belgie (PhDr. Marie Těšitelová, DrSc.)

Oběžníky ÚJČ
Pokyny SÚRPMO
Pracovní návody, pokyny
Pracovní řád
Vyhláška č. 100
Zápisy z porad
Zápisy ze schůzí
Závazky
Zpráva o činnosti oddělení matematické lingvistiky
Zpráva o činnosti Ústavu pro jazyk český

Administrativní styl: písemné texty

Hospodaření s domovním bytovým majetkem
Kolektivní smlouvy – TIBA
Materiál – TIBA
Metodické pokyny
Národní pojištění 12/1977

Administrativní styl: mluvené texty

Hlášení v metru
Hlášení v obchodním domě
Přehled rozhlasových pořadů
Zelená vlna
Zprávy o počasí