Zpravodaj Československého sdružení uživatelů TeXu

Karel Pala

Počítačový fond češtiny

Zpravodaj Československého sdružení uživatelů TeXu, Vol. 1 (1991), No. 4, 19–25

Persistent URL: http://dml.cz/dmlcz/148817

vsunout příslušný blok mezi `\begin{verbatim}` a `\end{verbatim}`). Dále je v souboru `verbatim.sty` definováno okolí `comment`, které má ten efekt, že všechen text mezi `\begin{comment}` a `\end{comment}` je při zpracování TEXem ignorován, Stylový soubor `theorem.sty` poskytuje mnohem více možností pro definování okolí typu **Věta, Tvrzení, Definice** atd. Tyto soubory (`theorem.sty` a `verbatim.sty`) je možné používat samostatně a to dokonce v rámci standardního LATEXu. Původně tyto soubory byly skutečně napsány jako stylové soubory pro standardní LATEX (ze stylových souborů autorů Mittelbacha a Schöpfa je u nás asi nejznámější styl `multicolumn`, který umožňuje paní vícesloupcového textu). Soubor `theorem.sty` bude automaticky volán při použití `amsart.sty` nebo `amsbook.sty`.

Už při používání standardního LATEXu se můžeme setkat s potížemi s pamětí např. při vícenásobném použití příkazu `\bezier` na jedné stránce nebo při mnohonásobném použití příkazu `\label` (návěstí pro křížové reference) nebo u delších textů dokonce jenom proto, že je příliš mnoho příkazů `\section` apod. V $\mathcal{AMS}$-LATEXu, který kombinuje LATEX s mnoha příkazy $\mathcal{AMS}$-TEXu, se dá očekávat, že tyto potíže budou ještě větší. Skutečně je tomu tak a pravda je bohužel taková, že používat $\mathcal{AMS}$-LATEX s běžnými verzemi TEXu pro PC (což je u nás stále asi nejběžnější) může být poněkud obtížné. Každopádně se většinou budeme pohybovat na pokraji omezení paměti. Potíže bude především činit hlavní paměť.

U běžných implementací TEXu bývá hlavní paměť omezena číslem přibližně 65 500. V uživatelské příručce $\mathcal{AMS}$-LATEXu jsou uvedeny např. tyto údaje, které se týkají zpracování nepříliš dlouhého článku, který obsahuje asi 50 vlastních definicí příkazů a asi 50 návěstí pro křížové reference. Je-li zpracován standardním LATEXem při použití stylu `article`, je využití paměti 51 376 (tyto údaje bývají na konci `.log` souboru). Je-li zpracován v $\mathcal{AMS}$-LATEXu ve stylu article a voláme-li navíc soubory `amsfonts.sty` a `amsbsy.sty`, je kupodivu využití paměti o něco menší — 51 059. Pokud v $\mathcal{AMS}$-LATEXu použijeme styl `article` a navíc volbu `amstex`, bude využití hlavní paměti již 63 506. Pokud dále necháme zpracovat tentýž článek ve stylu `amsart` (potom je automaticky volán `amstex.sty` a `theorem.sty`), bude využití hlavní paměti 65 445, což je již skutečně velice blízko maxima.

Zdá se, že používání $\mathcal{AMS}$-LATEXu vyžaduje větší implementace TEXu. Neměl jsem bohužel zatím možnost vyzkoušet $\mathcal{AMS}$-LATEX např. s BigTEXem (který je k dispozici v rámci emTEXu) nebo s nějakou jinou větší verzí TEXu.

(*Miroslav Dont*)
e-mail: `dont@cseara.bitnet`

Počítačový fond češtiny

Úvod

V tomto textu bychom rádi seznámili členy CSTUG s projektem Počítačového fondu češtiny, který si klade za cíl počítačové zpracování české slovní zásoby a vytvoření rozsáhlé lexikální databáze češtiny, jež bude sloužit jako východisko pro tvorbu českých slovníků všeho druhu.

I když je to snad všeobecně známo, pokládáme za potřebné připomenout, že významné slovníky hlavních evropských jazyků (angličtiny, francouzštiny, němčiny, italštiny, španělštiny a také ostatních) jsou v současnosti vytvářeny pomocí počítačů a také existují v tzv. *počítačově čitelné podobě* jako tzv. „machine readable dictionaries“. Jako příklady lze uvést Longman Dictionary of Contemporary English (LDOCE), Collins COBUILD English Language Dictionary, Oxford English Dictionary (OED), Oxford Advanced Learner's Dictionary of Current English (OALDCE), Merriam-Webster Seventh New Collegiate Dictionary a některé další. Počítačově čitelné slovníky slouží pak jako východisko pro tvorbu rozsáhlých *lexikálních databází*, které se stávají zdroji pro různé typy slovníků, např. pro dvojjazyčné či jinak specializované slovníky, jako jsou třeba slovníky pro počítačové zpracování přirozeného jazyka (Natural Language Processing) a umělou inteligenci (Artificial Intelligence).

Předpoklady

Lze konstatovat, že při formulování lexikografického projektu pro češtinu bylo již dosaženo celkové shody v několika základních oblastech, které mohou být zajímavé i pro členy CSTUG a uživatele systému TEX.

1. **Standardizace**
   Standardizace spočívá ve vytvoření co možná nejjednotnějších způsobů počítačového zpracování češtiny.
   1. Všechny podstatné problémy v oblasti kódů i editorů lze vyřešit připojením k celosvětovému projektu TEI (TEXT ENCODING INITIATIVE) a použitím v něm navržené techniky SGML (STANDARD GENERALIZED MARKUP LANGUAGE) (C. M. Sperberg-McQaueen & Lou Burnard, 1990).
   2. Pro vlastní práci na projektu však pokládáme za nezbytné, aby výchozím standardem se pro zainteresovaná pracoviště stal kód LATIN II, který navazuje na ostatní národní (západoevropské) kódy a je východiskem pro novou čs. normu. Skutečnost, že se užívá více různých kódů (s výraznou převahou kódu M. a J. Kamenických) lze bez větších obtíží řešit jednotným souborem konverzních programů, jenž by byl k dispozici všem zúčastněným pracovištím a pravděpodobně také členům CSTUG, pokud o to projeví zájem.
   3. standardní techniky zpracování typografických textových souborů (konverzní programy), které budou východiskem pro vytvoření reprezentativního korpusu českých slov (v první fázi cca 5–7 mil. slovních forem, v dalších fázích pak 20 a více mil. slovních tvarů). V tomto bodě se obracíme k členům CSTUG s výzvou ke spolupráci, která může přinést prospěch všem zúčastněným.
2. **Právní vztahy**
   Český korpus se bude vytvářet zpracováním typografických textových souborů, proto je nezbytné navázat kontakty s nakladatelstvími a tiskárnami a s pomocí právníků formulovat smlouvy, které umožní zpracovávat typografické soubory v souladu s platnými zákony o ochraně autorských práv. Návrh takové smlouvy opírající se o zahraniční zvyklosti je již k dispozici.[3]

[3] Lze konstatovat, že nedávno proběhla úspěšná jednání mezi členy Skupiny a redakcemi LIDOVÝCH NOVIN, MLADÉ FRONTY DNES o poskytování textových souborů vhodných pro vytváření korpusu češtiny. Navazují se kontakty s tiskárnami: Po-

Jádro projektu

Za hlavní cíl českého lexikografického projektu pokládáme vybudování **počítačového fondu češtiny**, který bude tvořen **lexikální databází současné spisovné češtiny** a **dílčími databázemi** zahrnujícími běžně mluvený jazyk, české a moravské dialekty a také starou a starší češtinu. Pro budování triády tvořené databází současné češtiny a dvěma dalšími databázemi se s výhodou použije jednotného programového vybavení a pokud možno vzájemně kompatibilních metodologických postupů.

Současná spisovná čeština

Pro lexikografické zpracování současného českého jazyka tedy předpokládáme tři základní úrovně:

1. úroveň **datové báze současných českých textů**, která by měla pokrýt slovní zásobu moderní češtiny a sloužit jako základ pro postupně budovaný reprezentativní korpus češtiny čítající kolem 20 miliónů slovních forem (pro srovnání s angličtinou viz např. Electric Word, May/June 1990, kde se počítá s korpusem v rozsahu 100 mil. slovních forem). Korpus bude postupně vznikat konverzí typografických textových souborů získávaných .z tiskáren a nakladatelství na ASCII soubory a vhodnou úpravou korpusů již existujících (např. ve VÚMSU).
2. úroveň **datové báze typu konkordance** obsahující hesla s kontexty a frekvenčními údaji a navazující na retrográdní slovník vytvořený v ÚJČ a rovněž i na soubor sémat vytvořený v ÚVTEI (Smetáček). Tato hesla mohou mít charakter „polotovarů“ a nemusí být zpočátku nijak homogenní.
3. úroveň **datové báze konkrétních slovníkových hesel** klasického typu (např. jako v SSJČ) nebo nového mezinárodního typu, která se předpokládají pro nové slovníky vytvářené v rámci evropského sdružení Language Industries. Tato báze by měla vzniknout buď sejmutím SSJČ pomocí scanneru[4] , nebo přepsáním SSJČ do počítače.
   - K tomu přistupuje pozitivní skutečnost, že *Slovník spisovné češtiny pro školu a veřejnost* (SSČ, ACADEMIA, Praha, 1987) existuje již v počítačově čitelná podobě a může tedy posloužit jako východisko pro další práci.
   - Pro práci s datovou bází slovníkových hesel není zatím momentálně k dispozici vhodný softwarový prostředek, tj. *specializovaný databázový systém*, který by byl okamžitě a bez výhrad použitelný pro budování české lexikální databáze.
4. Schůdné řešení pro všechny uvedené úrovně poskytuje výše citovaný projekt TEI (Text Encoding Initiative, C. M. Sperberg-McQueen & Lou Burnard, 1990), který umožňuje ukládat české texty, konkordance i slovníková hesla v textové podobě do běžných textových souborů a vnitřní hierarchie uvnitř hesel vyznačovat technikou SGML (Standard Generalized Markup Language).

---

lygrafia (Svoboda) a Spektrum (Tisk) a též se jedná o poskytnutí již existujících korpusů, např. ve Výzkumném ústavu matematických strojů.

[4] Počítačoví odborníci vyslovují však o schůdnosti této cesty značné pochybnosti vzhledem k velké typografické složitosti slovníkových textů. Pro daný účel pokládají za nejefektivnější slovník přepsat, ev. na smlouvu přesázet.

Uvedené datové báze poslouží jako zdroje pro vytvoření **lexikální databáze současné češtiny**, která pokryje celou slovní zásobu moderní češtiny a bude představovat strukturovanou zásobárnu údajů o jednotlivých vlastnostech českých lexikálních jednotek, a proto musí přirozené obsahovat podstatně více informace než kterýkoli existující český slovník. Z dobře koncipované a vybudované lexikální databáze bude pak možno zadáním vhodných kombinací třídicích kritérií vytvářet různé typy slovníků, s to i značně specializovaných.

Běžně mluvená čeština a české a moravské dialekty

Databáze běžně mluvené češtiny a českých a moravských dialektů by měly mít podobnou strukturu jako databáze současné češtiny a měla by také vznikat z podobných datových bází, tj. z báze mluvené češtiny, z báze nářečních textů, nářeční konkordance a odpovídajících slovníkových hesel.

Starší čeština

Lexikální databáze starší češtiny by měla zahrnovat jak staročeskou slovní zásobu, jak je obsažena v existujících staročeských slovnících, tj. např. ve *Staročeském slovníku*, ACADEMIA, Praha, díl 1–4, 1903–1984 a *Malém staročeském slovníku*, SPN, Praha, 1978, tak i starší češtinu z období baroka a obrození, což jsou úseky dosud nezpracované.

I zde se jako nejschůdnější postup jeví budování tří výše uvedených typů datových bází, tedy textové, konkordační a slovníkové.

Biblická datová báze

V současné kulturní a politické situací je zřejmé, že daný projekt bude jako rvou přirozenou část zahrnovat i *českou biblickou lexikální databázi* vycházející z nového ekumenického překladu Bible do češtiny. Český biblický korpus existuje již v počítačově čitelné podobě získané konverzí typografických souborů. Bez začlenění biblické datové báze do počítačového fondu češtiny by naše kulturní dědictví v oblasti české slovní zásoby nebylo zachyceno úplným a vyčerpávajícím způsobem. Jde ostatně o starý dluh v naší lexikografické tradici. Překladatelská skupina církve českobratrské evangelické již vyjádřila své pozitivní stanovisko.

Doplňující datové báze

Kromě uvedené triády obsahující *současnou češtinu, starší češtinu a české a moravské dialekty* pokládáme za potřebné začít systematicky vytvářet další tři nebo čtyři datové báze, které budou vhodným způsobem propojeny a mohou sloužit při vytváření lexikální databáze současné češtiny a také představovat zásobárny dat pro další výzkum.

Patří mezi ně:

1. **terminologická lingvistická databáze** obsahující údaje o lingvistických termínech získaných z jazykovědné literatury a periodik. Bylo by velmi žádoucí, aby

na vytváření této databáze se podílela všechna bohemistická, pracoviště v České republice.

2. **slovník českých kořenů a kmenů** a na něm založený derivační generátor a morfologický analyzátor. Je vcelku zřejmé, že tyto slovníky a programy mohou při vhodné implementaci fungovat jednak jako moduly uvnitř systémů pro zpracování přirozeného jazyka a jednak jako zdroje dat pro vlastní lexikální databázi. Díky své speciální povaze by však měly existovat i jako autonomní jednotky (moduly).
3. **úplný rejstřík (index)** všech slov a slovních forem obsažených ve všech právě uvedených databázích.
4. **úplný rejstřík (index)** všech slov a slovních forem obsažených ve všech právě uvedených databázích doplněný o veškeré **frekvenční údaje**. Tento rejstřík představuje východisko pro úplný **frekvenční slovník** češtiny.

Počítačový fond češtiny je koncipován jako plně *univerzální*, a má tedy umožňovat jak všestranný teoretický výzkum, tak i nejrůznější aplikace. Konkrétněji tedy půjde o možnosti konečně adekvátněji studovat úzus, o textovou analýzu, tvorbu nových gramatik a učebnic i konečně – v neposlední řadě – o prohloubené pokrytí paradigmatiky a syntagmatiky slova a jeho popis v nejrůznějších typech slovníků, především v *novém standardním slovníku českého jazyka.*

Hardware a software

Základní varianta

1. pro ukládání a zpracování textů a slovníků je potřeba počítat s jedním centrálním pracovištěm vybaveným hlavním počítačem typu RS6000 (IBM), SUN 4 nebo MIPS 3000 s operačním systémem UNIX, s vnitřní pamětí od 8 MB a diskovou pamětí od 1 GB výše (včetně standardního vybavení) (50 000 DM)
2. celé lexikografické pracoviště je potřeba vybavit lokální sítí ETHERNET tvořenou 10–12 počítači PC AT 386 SX (s pevnými disky 80 MB) (36 000 DM)
3. pro celé pracoviště je nezbytná výkonná laserová tiskárnou se systémem PostScript (20 000 DM)
4. lexikografické pracoviště se neobejde bez standardního kancelářského vybavení včetně 1–2 výkonných kopírovací přístrojů (jeden s formátem A3) plus dvě pevné telefonní linky a fax (12 000 DM)
5. pro archivování korpusů, konkordancí, slovníků a vytvořených lexikálních databází musí být lexikografické pracoviště vybaveno přepisovatelnými magnetooptickými paměťmi typu WORM a také jednotkami CD ROM (15 000 DM).

   V oblasti software je potřeba počítat s následujícím vybavením:

1. soubor programů tvořících tzv. LEXIKOGRAFICKOU PRACOVNÍ STANICI. Jsou to programy, které umožňují:
   - sestavovat z existujícího korpusu konkordance
   - získávat synonymické řady a popřípadě i hierarchie
   - poskytovat lexikografovi veškeré údaje potřebné při vytváření slovníkového hesla, patří k nim programové systémy jako WORDCRUNCHER, OXFORD CONCORDANCE PROGRAM (OCP) a/nebo LEXICOGRAPHER WORKSTATION (Calzolari, Picchi, 1990) aj.,

2. velký systém pro hlavní počítač a program PAT (nahrazující databázi a vyhledávací programy pro texty a slovníky) plus další programy získané v rámci výše citovaného projektu TEI.
3. OPERAČNÍ SYSTÉMY:
   1. operační systém MS–DOS pro práci na jednotlivých počítačích (uvnitř lokální sítě),
   2. operační systém UNIX pro hlavní počítač sítě.
4. vybrané TEXTOVÉ EDITORY a TEXTOVÉ PROCESORY pro práci s českými textovými soubory na jednotlivých počítačích, např. WORDPERFECT nebo CSED, T602, UNIX EMACS aj.
5. DATABÁZOVÉ SYSTÉMY – v počáteční fázi nebude potřeba pracovat s žádnými speciálními systémy (např. DBASE, FOXBASE či později SEZAM pod systémem MS–DOS), pod systémem UNIX připadá v úvahu systém INFORMIX.
6. programy pro TRANSFORMACI typografických souborů na textové, z nichž bude tvořen korpus.
7. programy pro ANALÝZU a VYHLEDÁVÁNÍ (systémy typu PAT) a ZPRACOVÁNÍ slovníkových hesel v počítačově čitelných slovnících.
8. programy pro MORFOLOGICKOU ANALÝZU (lemmatizátory) a SYNTAKTICKOU ANALÝZU češtiny.
9. vhodný publikační (DTP – desk top publishing) systém. Za takový systém pokládáme TEX, který plně vyhovuje předpokládaným potřebám projektu a patří do *public domain*).

Celkový odhad finančních nákladů na projekt činí 3,0–3,5 mil. Kčs.

Předpokládané výstupy a jejich využití

Počítačový fond češtiny bude poskytovat řadu výsledků vhodných pro využití v oblasti typografie, v redakcích, nakladatelstvích a všech institucích, v nichž jde o zpracování českých textů ve větším měřítku:

1. **Data**
   - reprezentativní korpus češtiny vytvořený použitím techniky SGML (viz výše)
   - různé typy konkordancí budované rovněž s použitím značek v rámci techniky SGML
   - datové báze a posléze i slovníky v počítačově čitelné podobě
2. **Procedury a programy**
   - Na základě uvedených dat lze získávat specializované a pokud možno úplné podklady pro budování:
   - dokonalejších automatických korektorů schopných korigovat nejen překlepy, ale i chyby gramatické (např. shodu a chybné vazby) a stylistické
   - co nejúplnějších programů pro české dělení slov
   - programové moduly umožňující komunikaci s počítačem v přirozeném jazyce (komunikace s datovými bázemi), automatické gramatiky pro strojový překlad apod.
3. **Slovníky a gramatiky**
   - Vytvořená data umožní také připravovat různé typy slovníků v textové i strojově čitelné podobě, orientované na různé skupiny uživatelů

– nové gramatiky založené na kvalitních datech a dostatečně úplném materiálu
– dokonalejší učebnice mateřského jazyka pro jednotlivé typy škol zachycující co nejpřesněji současnou normu naší mateřštiny
– kvalitní příručky a učebnice pro výuku češtiny jako cizího jazyka.

Odbornými garanty projektu **Počítačového fondu češtiny** jsou pracovníci vysokých škol a Ústavu pro jazyk český ČSAV, konkrétně doc. dr. František Cermák, DrSc., FF UK a ÚJČ ČSAV Praha, prof. dr. Petr Sgall, DrSc., FF UK Praha, dr. Karel Pala, CSc., FF MU Brno, doc. dr. Milada Hirschová, CSc., FF UP Olomouc, RNDr. Jan Králík, ÚJČ ČSAV Praha, dr. Eva Hajičová, DrSc., MFF UK Praha, doc. dr. Karel Kučera, CSc., FF UK Praha, RNDr. Jan Hajič, MFF UK Praha, dr. Věra Schmiedtová, ÚJČ ČSAV Praha.

V současné situaci se však dost dobře nelze spoléhat na to, že by ČSAV poskytla finanční prostředky pro projekt v dostatečné míře, proto se obracíme na domácí i zahraniční sponzory s žádostí o finanční podporu **Počítačového fondu češtiny**. Věříme, že i členové CSTUG projeví plné pochopení pro naše úsilí. (*Příspěvky lze posílat na bankovní účet České státní spořitelny 0800–2044345–018, konstantní symbol 0558*.)

(*Dr. Karel Pala, CSc.*)

MusicTEX—Sazba notových partů pomocí TEXu

Úvod aneb jak jsem k sazbě not přinucen byl

Letos na začátku června jsem se zúčastnil celostátní konference z teorie grafů. Setkal jsem se zde s problémy, které trápí především nečtenáře TEX bulletinu – a totiž jak a čím psát matematické články. Jako všichni, kteří propadli TEXu, jsem reagoval i já – hodinovou přednáškou na téma TEX. A abych všechny přesvědčil o nedostižných kvalitách TEXu, tvrdil jsem, že TEX umí „prostě všechno“. Na dotaz, zda umí sázet i noty jsem odpověděl svým vlastním citátem: „TEX umí vše, jenom se musí vědět jak“. Na toto zareagovali mí posluchači a tak jsem dostal za úkol vysázet TEXem „Hymnu teorie grafů“ (jedna z ukázek).

Nastala doba shánění informací, kdo, kdy a s jakým úspěchem se věnoval sázená not. Objevil jsem jediný balil: maker, nazvaný svým autorem D. Taupinem MusicTeX. S ním se mi po menší námaze podařilo vysázet už zmíněnou skladbu a tak jsem souhlasil s návrhem podělit se s vámi o svoje zkušenosti. Jsou to zkušenosti člověka naprosto neznalého notového zápisu, bez jakéhokoli hudebního sluchu. Toto mějte, prosím Vás, na paměti při posuzování mého článku.

K čemu lze použít MusicTeX

MusicTeX je „*public–domain*“ balík maker, určených k sazbě polyfonní a instrumentální hudby. Pomocí něj jste schopní vysázet šest nástrojů (přičemž hlas se počítá za dva nástroje - první pro hudbu, druhý pro text). Kódování pro TEX je