Štruktúra dát v Slovenskom národnom korpuse a ich vonkajšia anotácia

Radovan Garabík: Slovenský národný korpus, oddelenie Jazykovedného ústavu Ľ. Štúra SAV

Príspevok odznel na konferencii Slovenčina na začiatku 21. storočia, Prešov 7.3.2003.

Abstract
The paper "Data structure in Slovak National Corpus and its external annotation" presents proposed system for data storage and annotation in Slovak National Corpus. Taking into account the desirability to ensure readability of the data in future times, the system is targeted towards platform independent, easily decipherable annotation format. The data in corpus are organised into hierarchical structure divided into four different levels. Archive contains original documents without any other processing, Bank contains documents converted into common text format, Corpusoid adds linguistic annotation to the data, and finally the texts are converted into data suitable for corpus manager. Further we discuss the proposed common text format for documents in the Bank, as well as the structure of style-genre annotation of documents.

Výhľadovo si treba uvedomiť, že existuje pomerne veľká pravdepodobnosť, že v budúcnosti (perspektívne aj v značne vzdialenej) bude potrebné sa k dátam uloženým v korpuse vrátiť a ďalej s nimi pracovať. Preto sa snažíme zvoliť taký formát ukladania dát, ktorý je čo najmenej závislý na použitej počítačovej architektúre a dostupných programových prostriedkoch, s perspektívou udržania čitateľnosti podľa možnosti v priebehu nasledujúcich desaťročí. Preto bol pri návrhu štruktúry korpusu kladený obzvláštny dôraz na formálny návrh ukladania údajov.

Korpus je rozdelený na viac úrovní, v každej úrovni sa uchovávajú dáta v inom stupni spracovania.

Hierarchia sa skladá z nasledujúcich stupňov:

- archív
- banka
- corpusoid
- dáta pre korpusový manažér

Archív

Archív zdrojových textov je tvorený súbormi, ktoré sú priamo získané z externých zdrojov. Každý takýto text, získaný z jedného zdroja v jednom časovom okamihu, tvorí jednu základnú nedeliteľnú jednotku archívu, nazývanú pre naše účely „dokument". Fyzicky môže byť dokument tvorený viacerými súbormi, niekedy aj zložitou adresárovou štruktúrou. Tieto súbory sú v zdrojovom tvare bez akejkoľvek konverzie alebo úpravy a sú takto aj uchovávané. Napríklad dokumenty získané skenovaním sa ukladajú ako obrázky, ak to dovolia technické možnosti.
Archivované texty sú nedotknuteľné, pôvodné tlačové a pravopisné chyby sa neopravujú.

Banka

Banka obsahuje texty v spoločnom formáte inšpirovanom jazykom XML. V textoch v banke sú už odstránené všetky nadbytočné informácie, napríklad typografické údaje neobsahujúce informáciu o logickej štruktúre textu, vzorce,

tabuľky, obrázky a podobne. V texte sú vyznačené časti písané iným fontom a typom písma, ak sú použité na zvýraznenie alebo iné upozornenie na odlišnú funkciu v texte. Banka neobsahuje žiadne dodatočné (lingvistické) informácie, tie budú pridané až v ďalšej úrovni. Nad bankou je naplánované implementovať fulltextové vyhľadávanie, z licenčných dôvodov zrejme len pre vnútornú potrebu Slovenského národného korpusu.

Základnú nedeliteľnú jednotku banky nazývame „text“, pričom existuje jednoznačná väzba medzi textom v banke a dokumentom v archíve. V prípade, že dokument archívu sa skladá z viacerých bibliograficky identifikovateľných jednotiek, snažíme sa tento dokument rozdeliť na viacej textov, každý z nich jednoznačne identifikovateľný. Avšak aj v prípade že dokument je tvorený jedinou ucelenou bibliograficky určenou množinou dát, môže byť rozdelený na viacero textov, ak sú napríklad od rôznych autorov alebo rôzneho žánru, ako typický príklad uveďme knihu a doslov ku knihe.

Texty v banke sa občas môžu meniť, najmä v prípade opravovania chýb vzniknutých konverziou z dokumentov archívu.

Corpusoid

Corpusoid kopíruje svojou štruktúrou banku. Texty v corpusoide sú vnútorne anotované, väčšinou automaticky použitím externých programových nástrojov. Ručná vnútorná anotácia je osobitne označená. Predpokladá sa, že tieto texty budú často regenerované zo zdrojových textov banky.

Dáta pre korpusový manažér

Dáta, vhodné na spracovanie korpusovým manažérom automaticky generujeme z corpusoidu. Pri tejto konverzii môže prichádzať k strate informácií, pretože korpusový manažér nemusí podporovať všetky typy informácií nachádzajúcich sa v corpusoide. Aké informácie budú prístupné pre používateľov, záleží na použitom korpusovom manažéri.

V Slovenskom národnom korpuse predpokladáme v prvej etape nasadenie korpusového manažéra Manatee (`http://www.textforge.cz`).

Externá anotácia

Anotácia je tvorená štruktúrou kľúč-príznak-hodnota. Hodnota je vždy zapísaná ako reťazec znakov ukončený koncom riadku, čo vylučuje viacriadkové názvy. Hodnota môže byť buď voľná (napr. meno autora), alebo vybraná z fixnej množiny povolených hodnôt (napr. žáner).

Tieto hodnoty majú špeciálny význam:

- prázdne miesto: neurčená hodnota, túto hodnotu sme neurčovali, prípadne ju môžeme určiť neskôr, ale apriori o nej nič nevieme povedať.
- *XXX*: neznáma hodnota. Nevieme určiť (napr. meno autora v novinovom článku).
- *YYY*: neurčiteľná hodnota. Z princípu sa nedá určiť alebo nemá význam (napr. pohlavie autora, ak ide o kolektívne dielo, pohlavie prekladateľa, ak nejde

o preložené dielo).

- *MIX*: v prípade hodnoty vyberanej z fixnej množiny je hodnota zmes iných.
- *MSC*: v prípade hodnoty vyberanej z fixnej množiny je hodnota, ktorá nie je v danej množine určená.
- *TTT*: neznáma hodnota ktorú treba určiť – Značka, že k danej anotácii sa nevyhnutne treba neskôr vrátiť a hodnotu doplniť.

Hodnoty môžu byť bližšie označené týmito príznakmi:

- *fuzzy*: označuje nejasnosť hodnoty. Spolu s hodnotou označuje, že je určená len približne (a na rozdiel od TTT nie je potrebné alebo neplánujeme sa k záznamu vrátiť a hodnotu spresniť).
- *auto*: označuje, že daná hodnota bola automaticky vygenerovaná na rozdiel od hodnôt, ktoré zadal človek.

Implementácia archívu

Každý dokument je umiestnený v samostatnom adresári. S ohľadom na obmedzenia filesystému a na lepšiu orientáciu sa používa štvorstupňová hierarchia adresárov podľa dátumu *Acqdate*. V adresári zodpovedajúcom danému dokumentu sa nachádza zdrojový dokument. Okrem toho sa tam nachádza súbor *.annotation*, obsahujúci externú anotáciu. Zdrojový dokument môže byť skomprimovaný štandardným komprimačným programom *gzip(1)*, v takom prípade bude mať k pôvodnému menu pripojenú príponu *.gz*. Iné komprimačné programy (napr *bzip2*) nepoužívame. V prípade, že je dokument tvorený zložitou adresárovou štruktúrou, môže byť skomprimovaný vo formáte *.tar.gz*.

Implementácia banky

Pri implementácii externej bibliografickej a štýlovo-žánrovej anotácie sme vychádzali zo štruktúry anotácie v Českom národnom korpuse, s niektorými zmenami reflektujúcimi skúsenosti kolegov z ČNK a špecifické podmienky SNK.

Všetky texty v banke a corpusoide sú bez výnimky v kódovaní UTF-8, normalizácia C. Konce riadkov používajú unixovskú konvenciu, t. j. ASCII znak 0x0A, '\n'. Snažíme sa zbytočne nezneužívať kontrolné znaky v unicode, pretože tieto nebývajú dostatočne dobre zobraziteľné v bežných textových editoroch. Namiesto toho zavedieme XML tagy. Text v banke by mal odrážať logickú štruktúru dokumentu, nie typografické formátovanie. Avšak v mnohých prípadoch je logická štruktúra tesne zviazaná s typografickou a sformátovať súbor tak, aby odrážal logickú štruktúru, často vyžaduje neprimerane veľa manuálnej práce. Preto text v banke zachytáva podľa možností pôvodnú typografickú štruktúru dokumentu.

V banke uchovávame informáciu o dvoch typoch koncov riadku, mäkkom konci bez významnej logickej hodnoty, a o tvrdom konci s významnou hodnotou (napr. konce veršov v básňach).

Z ďalších typografický údajov zachytávame konce odsekov (značka **<p>**), nejasnosti v texte (značka **<noise**, nadpisy (**<h1>**Nadpis**</h1>**), zvýraznený text (**<em**>Zvýraznený text</**em>**).

Vzorce (matematické, chemické a iné) nahradíme značkou **<equation>**, jednoduché vzorce, zápisy fyzikálnych veličín a chemické zlúčeniny a reakcie (ak nesú podstatnú lingvistickú informáciu, napríklad H_2O je dostatočne vžité vo

všeobecnom povedomí slovensky hovoriacej komunity a je chápané ako slovo) zapíšeme použitím príslušných unicodových symbolov.

Tabuľky nahradíme značkou **<table>>**), obrázky značkou **<picture>**, malé obrázky v texte značkou **<icon>**.

Rozlišujeme nasledujúce typy pomlčiek:

- U+002D HYPHEN-MINUS (-) používame v banke na zápis nejednoznačnej pomlčky, keď nevieme pri konverzii textu automaticky rozhodnúť o aký typ ide – napríklad ak bol v zdrojovom texte použitý iba jeden typ znaku aj pre spojovník, aj pre pomlčku.
- U+2010 HYPHEN (-) používame v banke na zápis spojovníka. Príklad: „Rakúsko-Uhorsko“.
- U+2014 EM DASH (–) používame na zápis pomlčky. Príklad: „Peniaze – radosť“.
- U+2212 MINUS SIGN (−) používame na zápis unárneho alebo binárneho matematického operátora mínus, ak je v zdrojovom texte tento operátor odlíšený. Ak nie, používame U+002D HYPHEN-MINUS (-).
- U+00AD SOFT HYPHEN je dosť nejednoznačne definovaný. Interne ho používame na zápis hyphenation point na mieste, kde prišlo naozaj k rozdeleniu slova (inými slovami, výlučne iba na konci riadku). Tento znak nebude mať dôvod sa objaviť v banke.
- U+2011 NON-BREAKING HYPHEN považujeme za ekvivalent U+2010 HYPHEN a v banke ho nepoužívame.

Príloha

Anotácia archívu

Anotácia archívu obsahuje tieto kľúče:

- *Name*: meno dokumentu.
- *Acqdate*: dátum získania dokumentu pre korpus.
- *Id*: identifikačný reťazec dokumentu. Dokument nemusí byť fyzicky umiestnený na serveri, môže sa nachádzať na externom médiu s označením identifikačného reťazca.
- *Source*: zdroj, odkiaľ bol získaný dokument, zvyčajne názov vydavateľstva alebo redakcie.
- *License*: licencia dokumentu – rozlišujeme tri úrovne licencie, len pre vnútorné použitie pracovníkmi JÚĽŠ SAV, licenciu umožňujúcu zaradenie dokumentu na CD nosič, a nakoniec licenciu umožňujúcu sprístupnenie prehľadávania dokumentu na Internete.
- *Comment*: ľubovoľný komentár určený pre človeka (nie pre počítačové spracovanie).

Anotácia banky

Anotácia banky obsahuje tieto kľúče:

- *Name*: názov textu.
- *Origname*: pôvodný názov textu, ak ide o preklad. Zapíname *fuzzy* príznak, ak je názov transliterovaný.
- *Author*: meno autora. Tak, ako je uvedené v zdrojoch k dokumentu, podľa normy pre bibliografické zápisy.
- *Origauthor*: originálne meno autora v pôvodnom jazyku. Z technických dôvodov zrejme zostane väčšinou prázdne, vypĺňať sa bude ak bolo meno autora uvedené v preklade chybne.
- *Translator*: meno prekladateľa. *YYY*, ak nejde o preložené dielo.
- *Translation*: určuje či bol text preložený.

trn	preklad
org	originálny text
ftr	voľne preložený, prerozprávaný text

- *ISBN*: isbn číslo (ak existuje).
- *ISSN*: issn číslo (ak existuje).
- *SourceId*: Id dokumentu archívu, z ktorého dokument banky pochádza.
- *Rhyme*: rýmovanie dokumentu.

nrh	nerýmovaný
rhy	rýmovaný
MIX	kombinovaný

- *Type*: typ textu, presnejšiu špecifikáciu typu textu určuje kľúč *Subtype.*

	Type		**Subtype**
poe	**poézia**		
		ver	báseň
		son	pieseň
dra	**dráma**		
		scr	scenár, prepis rozhlasového vysielania
pro	**próza**		
		nov	román
		col	súbor poviedok, jednotlivá poviedka
		fac	literatúra faktu
inf	**informatívny text**		
		pub	publicistika
		adv	reklamné materiály, reklama, inzeráty
prf	**odborný text**		
		sci	vedecká literatúra, články, odborné časopisy, učebnice pre VŠ
		pop	populárnovedecká literatúra, záujmové časopisy
		txb	učebnice pre stredné a základné školy
		enc	encyklopédie a podobné abecedne usporiadané diela
		adm	administratíva
		man	manuály, návody na obsluhu
liv	**živá komunikácia**		
		spk	hovorený
		wri	písomný

- *Genre*: žáner/odborná oblasť, *Subgenre* určuje presnejšiu špecifikáciu

	Genre		**Subgenre**
img	**imaginatívne**		
		bel	beletria
		crm	detektívky, špionážne romány
		scf	sci-fi, fantasy
		jun	pre deti a mládež, rozprávky
int	**interstage**		
		trv	cestopisy
		mem	(auto)biografie, memoáre
		chr	kroniky
		let	listy
		ess	eseje
pol	**politika**		
ars	**umelecké vedy**		
		mus	hudba, opera, opereta, balet
		cin	film
		arc	architektúra
		art	výtvarné umenie, fotografia, sochárstvo
		the	divadlo, divadelná veda a kritika
		lit	literatúra, literárna veda a kritika
hum	**humanitné vedy**		
		his	história, archeológia
		psy	psychológia
		edu	pedagogika
		soc	sociológia, komunikácia, médiá
		phi	filozofia
		inf	knihovníctvo a informačné zdroje
		pol	politológia
		lin	lingvistika
		eth	etnológia, etnografia
law	**právo**		
		jur	právo
nat	**prírodné vedy**		
		agr	poľnohospodárstvo
		med	medicína
		zoo	zoológia
		bot	botanika
		bio	biológia
		che	chémia
		mat	matematika
		ggr	geografia
		phy	fyzika
		met	meteorológia
		geo	geológia
		env	environmentalistika, ekológia
tec	**technika**		
		tra	doprava, spoje, telekomunikácie
		ene	energetika
		ind	priemysel
		com	informatika, počítače

	Genre		Subgenre
		bui	stavebníctvo
		sta	normalizácia, štandardizácia
ecn	**ekonómia, riadenie**		
		eco	ekonómia, bankovníctvo, obchod
		mng	manažment, riadenie
		mer	tovaroznalectvo, spotrebiteľská oblasť
blf	**náboženstvo, viera**		
		rel	náboženstvo, viera, sekty
		teo	teológia
		exc	nadprirodzeno, okultné vedy, mágia
lif	**life style**		
		hou	domácnosť (byt, záhrada, ručné práce)
		fsh	odievanie, móda
		spo	šport
		sct	spoločenský život
		amu	zábava, hra, koníčky, voľný čas
		min	etnické minority
		reg	región
ins	**interdisciplinárne vedy**		

- *Medium*: médium

lib	kniha
nws	noviny
jou	časopis
ste	skriptá
net	internet a iné (pre-internetové) siete
for	formulár
occ	príležitostné
npu	nepublikované texty, rukopisy, školské práce
tvf	televízia
rad	rozhlas

- *Authsex*: pohlavie autora

msc	muž
fem	žena

- *Lang*: jazyk diela, trojpísmenová skratka podľa ISO 639-2 (slovenčina *svk*). Prípadné inojazyčné súvislé časti textu budú označené v corpusoide. Kompletne inojazyčný text nebude mať dôvod vyskytnúť sa v SNK, kým sa nezačne budovať paralelný korpus.

- *Varieta*: variant jazyka, v ktorom je dokument. Väčšinou bude spisovná slovenčina.

std	spisovná slovenčina
nst	neštandartná slovenčina

- *Transsex*: pohlavie prekladateľa.

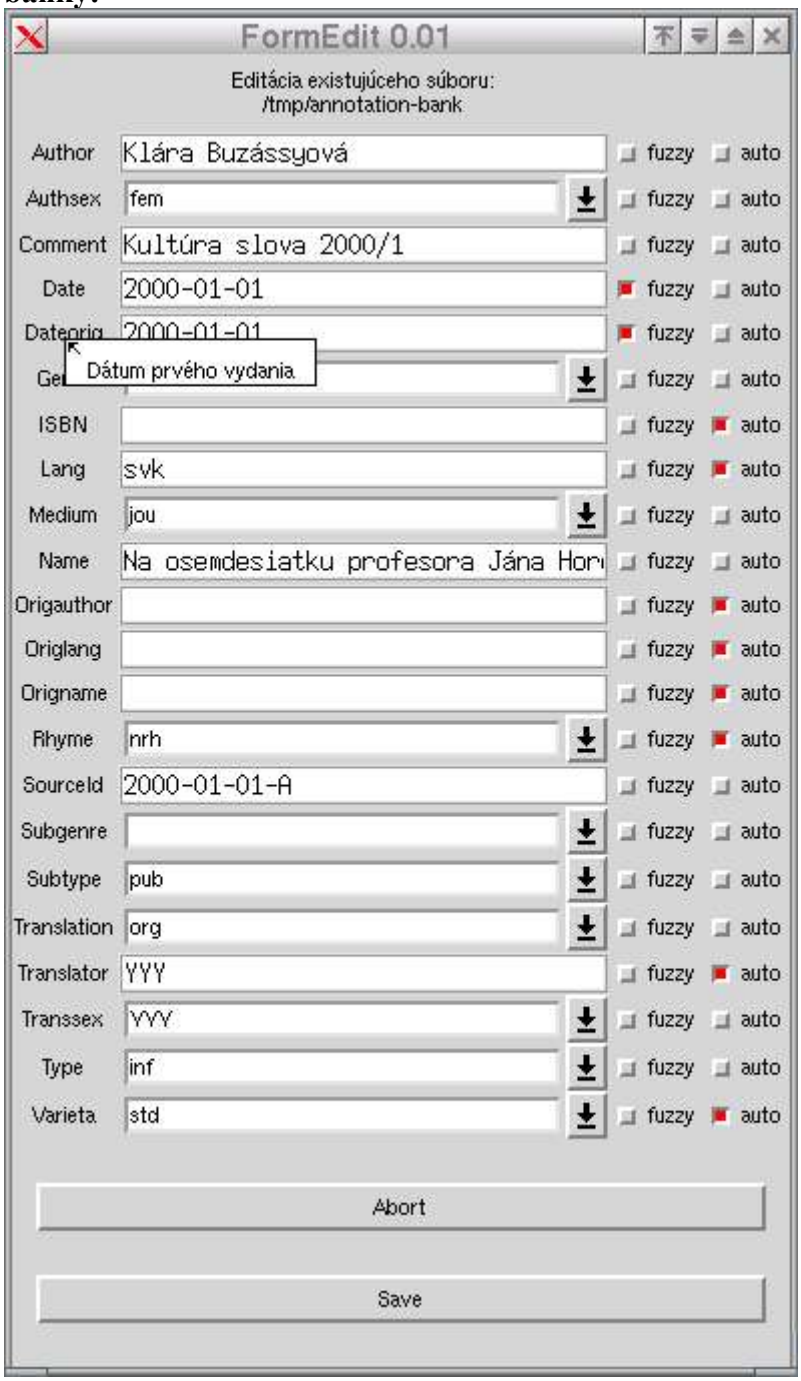
msc	muž
fem	žena

- *Origlang*: pôvodný jazyk diela.
- *Date*: dátum vydania.
- *Dateorig*: dátum prvého vydania (vzniku diela).
- *Comment*: komentár.

Príklad záznamu jedného dokumentu v archíve

```
Comment: Pravda
Source: Perex, a. s.
Name: Pravda 1991-04
Licence(auto): 1
Acqdate(auto): 2003-02-19
Id(auto): 2003-02-019-A
```

Prezeranie banky:



Príklad záznamu jedného dokumentu v banke

```
Name: Na osemdesiatku profesora Jána Horeckého
SourceId: 2000-01-01-A
Date(fuzzy): 2000-01-01
Dateorig(fuzzy): 2000-01-01
ISBN(auto):
Author: Klára Buzássyová
Authsex: fem
Origauthor(auto):
Origlang(auto):
Origname(auto):
Translator(auto): YYY
Transsex: YYY
Translation: org
Lang(auto): svk
Type: inf
Subtype: pub
Genre: int
Subgenre:
Medium: jou
Varieta(auto): std
Rhyme(auto): nrh
Comment: Kultúrne slovo 2000/1
```

Bibliografia

Leech, Geoffrey (2000): *Anotační systémy pro značkovaní korpusu*. In: Acta Universitatis Carolinæ – Philologica 3–4 – Studie z korpusové lingvistiky. Praha: Univerzita Karlova v Praze, Nakladatelství Karolinum, s. 185–197.

Kol. (2001): *Slovenčina a čeština v počítačovom spracovaní.* Bratislava: VEDA Vydavateľstvo Slovenskej akadémie vied.