Rok 2004, roč. 15, č. 3, s. 152-168

Český národní korpus – základní charakteristika a širší souvislosti

František Čermák, Věra Schmiedtová
Ústav českého národního korpusu FF UK
vera.schmiedtova@ff.cuni.cz

Poznámka: Autoři děkují pracovnímu týmu ČNK, který se podílel na přípravě tohoto článku svými připomínkami, podněty a návrhy.

Úvod

Český národní korpus (ČNK) je kontinuální projekt odrážející celosvětový trend v lingvistice, který spočívá v orientaci na lepší a spolehlivější informace, na jejichž základě lze dospívat k prohloubenému poznání o všech jazykových charakteristikách přirozeného jazyka. Pro češtinu je ČNK jedním z prvních produktů, který zároveň patří k jednomu z největších korpusů v Evropě. ČNK je veřejně přístupný na adrese http://ucnk.ff.cuni.cz, kde jsou k dispozici i další informace o projektu ČNK. Jako největší informační zdroj tohoto druhu u nás představuje již dnes významný národní kulturní statek.

Článek je rozdělen do dvou hlavních částí. V první, obecné části jsou definovány základní pojmy (korpusová lingvistika, korpus) a jsou přiblíženy nejdůležitější problémy spojené s budováním korpusů, v druhé části jsou podrobně popsány charakteristiky ČNK, jeho struktura, zpracování textů, možnosti využití a vyhledávání.

1 Korpus a korpusová lingvistika

1.1. Korpusová lingvistika

Korpusová lingvistika je disciplína lingvistiky zkoumající jazyk pomocí elektronických jazykových korpusů a zabývající se i výstavbou těchto korpusů, jejich zpracováním a příslušnou metodologií. Jako obor se začala výrazněji vyhraňovat a rozvíjet teprve v posledních dvou desetiletích 20. století, a to v souvislosti s rychlým rozvojem výpočetní techniky, který umožnil vznik rozsáhlých souborů jazykových dat v elektronické podobě a nevyhnutelně vedl k interdisciplinární spolupráci lingvistů s dalšími obory, především matematikou a počítačovou vědou. Přestože některé malé korpusy vznikaly už dříve (první z nich, *Brown Corpus*, k jehož tvůrcům patřil lingvista českého původu Henry Kučera, byl dokončen roku 1964 a zahrnoval jeden milion textových slov), teprve možnost počítačových operací s jazykovými daty o podstatně větších objemech (minimálně o jeden řád) vyvolala potřebu skutečně nového přístupu k této problematice.

Obecně řečeno, tento nový, korpusově-lingvistický přístup představuje především takové zkoumání textů, při němž se texty chápou jako rozsáhlé produkty jazykového systému a

schopností jejich tvůrců a skrze něž se dospívá k poznání obecnějších jazykových zákonitostí a pravidel. Korpusová lingvistika tedy není novou teorií jazyka: od jiných jazykovědných směrů se liší jen důsledným využíváním jazykových dat (k čemuž jí slouží počítače a speciálně vyvíjené softwarové nástroje), nově budovanou metodologií a velkým rozsahem těchto dat. Díky velkým korpusům (tj. korpusům o rozsahu stovek milionů textových slov, mezi něž se v roce 2000 zařadil i ČNK) je lingvista poprvé v historii zbaven nejistoty, zda nepracuje s příliš omezenou materiálovou základnou a zda pozorování a závěry, které z ní vycházejí, nejsou proto deformované. Korpusová data lze vzhledem k jejich mimořádnému rozsahu obvykle charakterizovat jako (1) typická, nenáhodná a věrná ve vztahu k tomu, jak lidé užívají jazyka; dále jako (2) aktuální, resp. skutečně odrážející svou dobu, (3) neselektivní a objektivní, (4) dostatečná a (5) s pomocí počítače snadno získatelná a rychle přístupná. Korpus vzniklý promyšleným sběrem textů ve velkých objemech tak vylučuje nejen malou typičnost dat a vliv náhody při jejich sběru, ale i omezenou aktuálnost či deformující selektivnost danou výběrem a přehlédnutími; především však odstraňuje nesmírnou pracnost tradičního manuálního získávání dat. Manuální korpusy v podobě kartoték a archivů excerpt, které existovaly už dávno před vznikem korpusů elektronických, měly většinou právě tyto nedostatky a nevýhody.

První etapa budování elektronických korpusů byla přirozeně omezena na korpusy malého rozsahu (zhruba jeden milion textových slov), jejichž prototypem byl již zmíněný *Brown Corpus*. S rostoucími možnostmi výpočetní techniky si však lingvisté začali uvědomovat nejen to, že kvalita dat a jejich informační hodnota je ve velkých korpusech podstatně větší, ale také to, že velkých korpusů lze využít i pro nové cíle v rámci nových disciplín. Průkopnickou zemí v budování velkých korpusů se stala Velká Británie, v níž se také jako v první zemi konstituovala nejvýznamnější z nových korpusových disciplín **– korpusová lexikografie** (dnes jsou v Británii už prakticky všechny nové slovníky angličtiny založeny na korpusových datech). V současné době jsou největšími britskými korpusy *Bank of English* (korpus, který přesáhl hranici 500 milionů slovních tvarů a kolem nějž v Birminghamu vzniklo péčí známého britského lingvisty Johna Sinclaira významné slovníkové nakladatelství *Cobuild*), dále reprezentativní *British National Corpus* (*BNC,* korpus o rozsahu zhruba 100 milionů slovních tvarů, obsahující i významnou složku mluvenou, vytvořený ve spolupráci řady institucí v 90. letech) a *International Corpus of English*, jehož cílem je mapovat všechny ve světě užívané varianty anglického jazyka. K významným korpusům jiných jazyků patří zejména tři korpusy němčiny (v Mannheimu, Berlíně a ve Stuttgartu), francouzský *Frantext* a korpusy dánštiny, italštiny a španělštiny; rozsáhlé korpusy však vznikají i v Maďarsku, Polsku, Slovinsku, Litvě a jinde. Celkové množství korpusů ve světě nebo i jen v Evropě lze dnes už těžko odhadovat; pro Evropu v každém případě platí, že je obtížné vůbec najít jazyk, který by korpus neměl, nebo pro který by nebyl korpus připravován.

1.2 Korpus

Korpusová lingvistika chápe základní prostředek svého výzkumu, **korpus**, jako nejlepší aproximaci, nejvěrnější vzorek skutečného jazyka i veškeré informace, které jazyk zprostředkovává, a vychází tak z přesvědčení, že lépe než prostřednictvím korpusu nelze dnes jazyk při studiu uchopit. Korpus se obvykle vymezuje jako strukturovaný, unifikovaný (a často též označkovaný) rozsáhlý soubor jazykových dat, který je elektronicky uložený i zpracovávaný; skládá se zpravidla z jednotlivých textů a jako celek si činí nárok na

reprezentativnost vzhledem k vytčenému cíli. Jde o vymezení běžně přijímané, avšak k jednotlivým jeho částem je třeba připojit několik poznámek.

Vysvětlení zasluhuje především klíčový pojem **reprezentativnost**, který je – jak je zřejmé z výše uvedeného vymezení – chápán pouze relativně, ve vztahu k cíli, pro který má být korpus využit, neboť povaha a obecnost takových cílů se může značně různit (cílem může být výzkum jazyka jednoho autora nebo žánru právě tak jako výzkum jednoho jazyka jako celku nebo i paralelní výzkum několika jazyků). V případě velkých korpusů, které jsou zpravidla budovány jako materiálové základny pro výzkum současného stavu celého národního jazyka, se v souvislosti s reprezentativností obvykle zdůrazňuje odpovídající zastoupení jednotlivých typů jazyka z hlediska *recepce*, tj. z hlediska míry, v jaké jsou mluvčími přijímány (vnímány, čteny). Recepce se v tomto pojetí nadřazuje *produkci* (míře, v jaké někteří lidé jednotlivé typy jazyka aktivně produkují), protože lidé v průměru mnohem méně píšou, než čtou (skutečně plodné autory lze najít pouze v některých profesích), a rovněž méně mluví, než poslouchají. Tato koncepce reprezentativnosti, propracovaná na základě několika průzkumů, se stala východiskem i pro vnitřní strukturaci ČNK (podrobněji viz 2.3). Pro úplnost je třeba dodat, že reprezentativnost se v mnoha korpusových projektech chápe podstatně volněji a váže se především na rozsah korpusu, bez výrazného zřetele k jeho vnitřní strukturaci (např. v rámci evropského projektu EAGLES (Expert Advisory Group on Language Engineering Standards)

se mluví – vedle množství možných specifických pojetí korpusu – o základním, neutrálním či bezpříznakovém typu korpusu, za jehož podstatné vlastnosti se považuje především *velký rozsah*, dále *autentičnost textů*, z nichž je vytvořen, a *spolehlivou dokumentaci* jazykových dat z hlediska jejich původu).

Okrajovou, spíše historicky orientovanou poznámku zasluhuje ta část výše uvedeného vymezení korpusu, která poukazuje k dnes již převládající praxi budovat velké **korpusy z celých textů** a zachovávat tak celistvost a úplnost informace v nich, která je jiná například na začátku, uvnitř a na konci textu. Starší korpusy naproti tomu zpravidla vycházely ze statisticky vybíraných textových vzorků o standardní délce (např. 2000 slov); jde o metodu, která sice snižuje nebezpečí jednostrannosti malého korpusu a zvyšuje jeho reprezentativnost, avšak její očividnou nevýhodou je to, že na základě takto koncipovaného korpusu lze hůře studovat jevy celotextové povahy a je zapotřebí mnohem více textů. Ze současných velkých korpusů je jako vzorkový vybudován zejména výše zmíněný *British National Corpus*.
V českém korpusu jsou základními zařazovanými jednotkami zásadně celé texty.

Specifický aspekt problematiky jazykových korpusů představuje **značkování**, které podstatně zvyšuje užitnost korpusů. Základním typem značkování je tzv. *vnější anotace*, která vychází vstříc praktickému požadavku vědět, z jakého typu textu, z kterého roku, od jakého autora apod. pocházejí jednotlivé citace v korpusu. Tato informace bývá často zajímavá i pro nelingvisty (a lze se pak takto ptát na některá klíčová slova a šířeji i dobové pojmy a jevy a jejich začátek, jako např. *tunelovat, mantinely* aj.). Tyto informace se zanášejí do textu a programy sloužící k práci s korpusem je pak (obvykle ve formě zkratky) uvádějí na okraji každého konkordančního řádku dokumentujícího výskyt hledaného slova, tvaru apod. Kvalitní programy na vytěžování korpusu nabízejí i možnost vyhledávat v něm podle specificky vybraných parametrů a zvolit si např. vytvoření konkordance jen z textů určitého roku, jen z textů napsaných ženami, z textů určitého žánru apod. (pro potřeby ČNK se takový program dolaďuje). Formát, v němž jsou potřebné informace ve většině dnes budovaných korpusů přidávány k textům, je standardizovaný (ČNK podobně jako další velké korpusy užívá mezinárodně uznávaného formátu *SGML* (Standard Generalized Markup Language) a využívá

zásad iniciativy *TEI* (Text Encoding Initiative), které jsou založeny na konsensu několika vlivných mezinárodních organizací a představují unifikovaný soubor instrukcí, jak kódovat texty a výsledky jejich analýzy prostřednictvím jazyka SGML (viz dále 2.4)).

Užitnost korpusu dále podstatně zvyšuje *vnitřní anotace*, vnášející (zpravidla automaticky nebo poloautomaticky, s využitím speciálních jazykových programů) do korpusu *strukturní informace* (informace o členění textů na kapitoly, odstavce, věty, slova) a *informace lingvistické*. Teoreticky by bylo možno korpus obohatit lingvistickými informacemi libovolného druhu a v libovolném množství, avšak u velkých korpusů – zvláště u výrazně flektivního jazyka, jakým je čeština – je lingvistická anotace nesmírně pracná a drahá, a proto se v praxi omezuje nejčastěji na *morfologické značkování* jednotlivých slovních tvarů (tzv. *tagování*), zahrnující i přiřazení *slovnědruhové charakteristiky* a *lemmatizaci*, tj. přiřazení základního, slovníkového tvaru (podrobněji viz 2.4.4). Značkování víceslovných jednotek (jak značkování složených gramatických jednotek, k jakým patří např. složené časy, tak značkování frazémů či víceslovných termínů), jehož potřeba se ukazuje jako stále naléhavější, stojí pro svou obtížnost dosud na samém počátku vývoje.

1.3 Typy korpusů

Základní členění je lingvistické a dělí korpusy jednak na psané a mluvené, jednak na synchronní a diachronní. Převažujícím typem korpusu je dnes **synchronní korpus psaný**, tj. korpus založený na současných psaných textech, jejichž analýza je nejpotřebnější a jež jsou zároveň (elektronicky) nejdostupnější. Protože vývoj jazyka je kontinuální a žádné přirozené časové hranice v něm nejsou, volí se rozsah synchronního korpusu zpravidla tak, že zahrnuje několik posledních desetiletí, s ohledem na nesnadno objektivizovatelnou aktuálnost a životnost textů nebo na významné vnější, zejména společenské přeměny (o časovém vymezení ČNK viz blíže 2.2).

Diachronní korpus naproti tomu pokrývá několik vývojových stadií daného jazyka, popřípadě celý jeho vývoj. Z povahy věci je zřejmé, že diachronní korpus se liší od korpusu synchronního řadou vnějších i vnitřních charakteristik, mimo jiné i odlišným pojetím reprezentativnosti, jež lze vztáhnout pouze k úhrnu dochovaných jazykových textů, který ovšem bývá zejména ve starších vývojových stadiích jazyků značně omezený a nevyrovnaný co do zastoupení jednotlivých druhů textů (obvykle silně převažují texty náboženské, legendické, veršované a jinak netypické z hlediska celkového jazykového úzu).

Reprezentativní **mluvený korpus** (synchronní, o diachronním není z pochopitelných důvodů možno vůbec uvažovat) je zatím stále spíše jen velmi nákladným deziderátem než realitou. Příčinou je především skutečnost, že investice a úsilí, které je nutno do vytvoření mluveného korpusu vložit, jsou mnohonásobně vyšší než u korpusů psaných a v současnosti leží zcela mimo reálné možnosti většiny korpusových projektů; pokud některé mluvené korpusy přesto omezeně vznikají, jsou jen malé a obecnější charakteristiky mluveného jazyka tedy spíše jen naznačují než plně dokumentují. Vedle investiční a časové náročnosti zůstává při tvorbě mluveného korpusu i nadále nemalým problémem už samo získání autentických vzorků mluveného jazyka, který svou spontánností a neformálností tvoří přirozený protiklad jazyka psaného (v podstatě jediná současná možnost – magnetofonové nahrávání – spontánnost i

neformálnost mluvených projevů většinou značně narušuje). Koncepce ČNK se nepřiklání k názoru, že mluvený korpus lze vybudovat primárně z veřejných projevů získaných z médií, neboť jde o jazyk přechodné povahy, s vysokým podílem nespontánnosti, formálnosti a připravenosti, který je ovlivněn jazykem psaným; často však jde jen o jazyk čtený. Obtíže spojené s tvorbou mluvených korpusů provázejí i snahy o budování synchronních **nářečních korpusů**, k nimž se však navíc přidávají komplikace s vypracováním adekvátního pojetí reprezentativnosti, které by v tomto případě mělo zahrnovat dosud jen zcela neuspokojivě řešený vztah dialektů k věkovým, sociálním, kulturním, teritoriálním a jiným faktorům.

Paralelní korpusy jsou korpusy dvou nebo více jazyků vytvářené z překladů, obsahují tedy vždy jednak texty originální, jednak jejich jinojazyčné mutace. Tyto korpusy nabývají v poslední době stále více na aktuálnosti. Už první zkušenosti naznačují, nakolik bohatší a různorodější jsou cizojazyčné ekvivalenty užité dobrými překladateli v porovnání s klasickými dvoujazyčnými slovníky, a dávají tušit, jak rozsáhlé obohacení těchto slovníků paralelní korpusy v budoucnu nepochybně přinesou. Zároveň se na prvních pokusech ukazuje, jak cenný je autentický překladový materiál pro výuku studentů, překladatelů apod. V Evropě je dnes k testovacím a komparačním účelům k dispozici např. CD s produktem projektu TELRI (Platónova *Ústava* v 17 jazycích) nebo paralelní korpus románu G. Orwella *1984* ve 23 jazycích, který vznikl v rámci projektu Multext-East.

Pro studenty a jejich výuku se začínají vytvářet i **studijní korpusy** (*learners' corpora*), složené zvláště z textů psaných studenty cizích jazyků. Přestože jde o aplikaci korpusů, která je teprve na samém počátku vývoje, ukazuje se, že hromadná analýza způsobů vyjadřování, odstínů i chyb ve studentských textech může v budoucnosti vést k významnému zkvalitnění výuky.

Čistě technickou povahu mají **cvičné** a **testovací korpusy**. Jsou to rozsahem omezené korpusy, v nichž bylo zpravidla plně provedeno a manuálně opraveno značkování (vnitřní anotace). Na základě dat v těchto korpusech se následně trénují a vylepšují různé anotační programy i lingvistické hypotézy (podrobněji viz 2.4.5).

Konečně existuje i nepřehledná řada čistě tematických korpusů vytvářených pro potřeby jediného oboru či odvětví.

2 Český národní korpus

2.1 Obecná charakteristika

Český národní korpus (ČNK) je kontinuální projekt, jehož produkty (jednotlivé konkrétní korpusy) mapují a monitorují různé podoby českého jazyka s cílem zpřístupnit uživatelům co nejbohatší zdroj jazykových dat a příslušné nástroje k jejich využívání. ČNK je cílevědomě budován tak, aby nabízel co největší možnosti a zároveň byl schopen uspokojit co nejširší potřeby badatelů i pedagogů, odborníků i studentů, lingvistů i nelingvistů. Protože jde o projekt akademický, nekomerční, je otevřen bez většího omezení všem seriózním zájemcům.

ČNK je vytvářen Ústavem Českého národního korpusu (ÚČNK) na Filozofické fakultě Univerzity Karlovy (FF UK). ÚČNK, vedený prof. Františkem Čermákem, byl založen roku 1994 na základě iniciativy řady jednotlivců z různých pracovišť, kteří začali už před lety

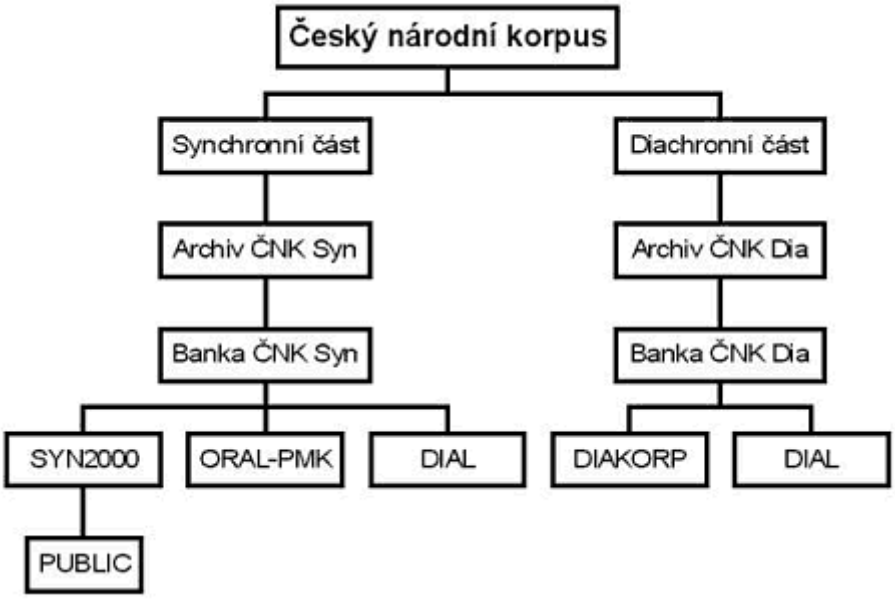
pociťovat naléhavou potřebu vybudovat velký korpus, jenž by byl dostatečnou materiálovou (datovou) základnou umožňující tvorbu nových, kvalitativně lepších slovníků češtiny, gramatik a dalších jazykových příruček. Samotný vznik ÚČNK je spojen především s ochotou zástupců FF UK, kteří projevili pochopení pro záměr vybudovat korpus českého jazyka a poskytli proto Ústavu institucionální základnu spolu s elementárním vybavením. Rozvoj činnosti ÚČNK byl pak rozhodující měrou podpořen získáním několika **grantů** Grantové agentury ČR, dále udělením grantu Ministerstva školství, mládeže a tělovýchovy a především nejnověji přiznáním institucionální podpory v rámci programu výzkumných záměrů téhož ministerstva. Významnou měrou podporují projekt ČNK také **sponzoři** (dříve Komerční banka, nyní zejména a dlouhodobě Nakladatelství Lidové noviny).

Na tvorbě ČNK se s ÚČNK podílí různým způsobem na smluvním základě několik domácích pracovišť, a to Ústav českého jazyka a teorie komunikace, Ústav bohemistických studií, Ústav teoretické a komputační lingvistiky (všechny FF UK), Ústav formální a aplikované lingvistiky (Matematicko-fyzikální fakulta UK), Ústav českého jazyka (Filozofická fakulta MU), Katedra informačních technologií (Fakulta informatiky MU), Ústav pro jazyk český AV ČR, Ústav pro českou a světovou literatury AV ČR, katedra počítačů fakulty elektrotechnické ČVUT a některé další. ÚČNK od svých počátků těsně spolupracuje rovněž s předními korpusovými pracovišti v zahraničí a účastní se mj. řady evropských projektů.

ÚČNK odpovídá za kontinuální budování a rozvoj ČNK i za související činnosti, zejména za výzkum, výuku a pěstování oboru korpusová lingvistika (v rámci matematické lingvistiky). Nedílnou součástí práce na ČNK je vytváření vlastní metodologie, neboť ze zahraničních korpusových projektů lze převzít jen některé zkušenosti obecnějšího rázu. V prvních etapách budování ČNK byla v tomto směru mimořádně naléhavá zejména problematika konverzí mezi různými formáty textových souborů a problematika způsobů anotace, která se musela opírat o lingvistickou analýzu a teorii aplikovanou za použití mezinárodních zásad TEI a jazyka SGML (viz 2.4). Průběžně však byla a je rozvíjena vlastní koncepce reprezentativnosti korpusu (viz 2.3) a rovněž průběžně je vytvářen a zdokonalován software pro účelné a efektivní vytěžování ČNK, zejména korpusový manažer (v současné době je používán GCQP; vyvíjen a ověřován je nový manažer Bonito), pro který bylo nutno řešit otázky optimálních filtrů a způsobů dotazování (v této oblasti bylo možno vedle vlastních přístupů využít mezinárodně osvědčených statistických parametrů a vzorců; další výzkum tu pochopitelně pokračuje a do jisté míry závisí i na objevujících se praktických potřebách uživatelů ČNK). Do budoucnosti zůstává významnou výzvou zejména propracování a rozvoj metodologie v oblasti variability jazyka a škálovité, odstupňované povahy informace.

2.2 Členění ČNK

Označení **Český národní korpus** je střešní název pro několik entit a složek v různém stadiu rozpracovanosti a vývoje, vytvářených z elektronických textů různé povahy, zaměření a rozsahu. Základní členění ČNK naznačuje následující schéma:



Znázorněné složky ČNK mají různou povahu a rozsah. Velikostí jim jednoznačně dominuje *synchronní psaný korpus* (SYN2000, rozsah přes 100 milionů slovních tvarů), ze kterého vychází korpus PUBLIC, veřejně přístupný na internetu (zhruba 30 milionů slovních tvarů). Podstatně menší je *diachronní psaný korpus* (DIAKORP, v současné době zhruba 1,75 milionů tvarů) a *synchronní mluvený korpus* (ORAL-PMK – Pražský mluvený korpus, asi 700 000 tvarů; BMK – Brněnský mluvený korpus, asi 500 000 tvarů); nářeční korpusy obojího typu (DIAL, synchronní a diachronní) mají dosud z praktických důvodů (viz výše 1.3) zatím spíše jen orientační a výhledovou povahu (celkem zahrnují pouze přes 100 000 tvarů). Jednotlivé korpusy jsou budovány zásadně z celých textů (tj. nikoli ze vzorků, srov. výše 1.2), respektive z celých dochovaných částí textů v případě některých děl ze starších vývojových fází češtiny.

V pozadí vlastních korpusů ČNK stojí nezbytné **archivy,** zejména *Archiv ČNKSyn* a *ČNKDia* (v nichž jsou uloženy všechny získané texty ve výchozí, tj. surové, nezkonvertované a neoznačkované podobě), a **banky,** zejména *Banka ČNKSyn* a *Banka ČNKDia* (v nichž se ukládají všechny texty v konečném formátu SGML; zveřejněnou a navenek známou částí *Banky ČNKSyn*, vybranou s ohledem na reprezentativnost, je mj. korpus SYN2000).

Korpusové složky ČNK jsou představeny v následujících odstavcích.

Synchronní korpus. Hranice mezi synchronním a diachronním korpusem

V jazyce se nenabízejí žádná jasná a nepochybná kritéria pro stanovení jeho časových rozmezí, a tedy ani pro stanovení hranic jednotlivých druhů korpusů. Časové dělení ČNK je také proto do jisté míry arbitrární a závislé na činitelích vnějších, historických. Pro minulost se tyto hranice volně odvozují od většího úhrnu změn v jazykovém systému. Jistým vodítkem pro oddělení současného jazyka od jazyka staršího (resp. od řady jeho různých starších podob) je však vždy to, jak se dnešní mluvčí k jednotlivým formám jazyka stavějí a co pociťují jako ještě živé a co už nikoliv. V podstatě z těchto hledisek byly také vymezeny spodní časové hranice v rámci ČNK:

1. V oblasti novinových a časopiseckých textů byl za začátek textového mapování a zařazování do *synchronního psaného korpusu* přijat rok 1990, a to vzhledem ke svému přelomovému charakteru (starší noviny, plné dobového ideologického newspeaku, mohou dnes už jen těžko reprezentovat skutečně současný jazyk, který se právě v publicistické oblasti nejrychleji mění).
2. Rok 1990 byl přijat jako počátek i v případě krásné literatury, avšak vzhledem k tomu, že literární texty se často znovu přetiskují a hojně se čtou i knihy některých starších autorů (ti v tomto smyslu patří do jazykové současnosti, protože ji svým vlivem spoluvytvářejí), bylo pro tuto oblast stanoveno několik dalších, vzájemně se doplňujících kritérií. Na jejich základě se do synchronního korpusu zařazují také: (1) v současnosti čtení starší autoři, kteří se narodili roku 1880 a později, a (2) knihy publikované od roku 1945, tj. od konce 2. světové války (tato díla jsou však v korpusu (podle jistého klíče) zastoupena řidčeji než texty publikované poprvé po roce 1989).
3. Z odborných textů jsou do synchronního korpusu zařazovány rovněž pouze ty, které vznikly po roce 1989, nebo byly znovu vydány.

Všechny texty, které leží za těmito časovými hranicemi, jsou řazeny do diachronního korpusu. Je však třeba připomenout, že dosud ještě ani zdaleka nejsou k dispozici všechny texty vhodné pro zařazení do ČNK (dostupnost textů je trvalým problémem) a že i u dostupných a daným časovým a jiným kritériím vyhovujících knih a textů bylo v případě SYN2000 nutno přikročit k proporcionálnímu výběru (např. beletrie je tu zastoupena jen cca 15 %, viz 2.3). Bylo by tudíž omylem domnívat se, že v synchronním korpusu či bance ČNK je v elektronické formě dostupná celá česká literatura daného období, popř. že v rámci projektu ČNK je převod tak rozsáhlého souboru dat proveditelný, jakkoliv by to bylo ideální. Korpus obsahující všechny současné texty a autory dnes neexistuje nikde ve světě a zřejmě nebude existovat ani v budoucnu. Tvůrci ČNK obecně velmi vítají spolupráci kohokoliv na tomto poli.

Horní časovou hranicí pro zařazení textů do synchronního psaného korpusu bylo časové rozmezí let 1998–1999, dané v podstatě dostupností textů (ty lze získávat vždy jen s určitým časovým odstupem, prodlouženým o čas, který si vyžádá jejich interní zpracování). Takto pokrytá časová etapa (1990–1999, s naznačenými přesahy do minulosti) je reprezentována v roce 2000, tj. v době dokončení první etapy ČNK, již zmíněným korpusem SYN2000 v rozsahu přibližně 100 milionů textových slov označkovaných vnějším i pokusným vnitřním, lingvistickým značkováním (blíže viz 2.4.2 a 2.4.3). Projekt ČNK ovšem pokračuje dále a na tuto první etapu a verzi navážou verze další a rozsáhlejší.

SYN2000 zahrnuje takové množství jazykového materiálu, jaké u nás dosud nikdy k operativnímu prohledávání a obecnému užití nebylo zpřístupněno; samotné texty (bez značkování) mají rozsah 1–2 GB a jejich prosté manuální prohlížení přesahuje lidské možnosti. Toto množství si lze při průměrné podobě a velikosti tištěných knižních stránek a při průměrné knize o 250 stranách tištěných na tenkém papíru představit jako 10 zaplněných metrů knihovních regálů. Pouhé přečtení celého synchronního korpusu by při poměrně rychlém tempu (150 slov za minutu, 8 hodin denně a 365 dní ročně) zabralo přes 4 roky. Podstatné však je, že příslušný software tento rozsah prohlédne a výsledek uživateli představí (podle složitosti dotazu) během několika sekund.

Diachronní korpus

Diachronní korpus ČNK (DIAKORP) je budován s cílem vytvořit elektronickou materiálovou základnu pro výzkum vývoje českého jazyka od prvních dochovaných souvislejších záznamů (2. polovina 13. století) zhruba do poloviny 20. století (s přesahem do konce 80. let 20. století v případě novinových a časopiseckých textů), tj. k hranicím synchronního korpusu. Do diachronního korpusu jsou zařazovány pouze dobově a útvarově autentické texty, tj. texty, u nichž lze s rozumnou mírou jistoty vyloučit, že do nich byly vneseny prvky pozdějšího jazykového stavu nebo jiného jazykového útvaru (k takovým neautentickým textům se s výjimkou kritických edic počítají všechny opisy a přetisky, které vznikly podstatně později než originál, nebo které původní text zjevně zkreslují nářečními a jinými prvky).

Postup budování diachronního korpusu je pomalý a obtížný, neboť většinu textů je třeba manuálně přepisovat nebo (v případě spolehlivých novodobých edic) skenovat a manuálně korigovat. Vzhledem k tomu, že elektronické prohledávání různorodých paleografických podob textů z odlišných období vývoje českého pravopisu je prakticky nezvládnutelné, vstupují texty z doby před rokem 1849 do diachronního korpusu v transkribované podobě. V současné době není v možnostech ČNK standardně připojovat k transkribovaným starším textům jejich transliterované podoby, avšak pro budoucnost se počítá s podstatně užším spojením transkripce s originálem, a to ve formě elektronického propojení korpusových transkribovaných textů s digitalizovanými obrazy jednotlivých stránek původních předloh. Toto řešení by mělo nejen umožnit badatelům detailně si ověřit jednotlivé případy transkripce, ale i podstatně rozšířit možnosti korpusového výzkumu na oblast vývoje grafiky a pravopisu; realizace tohoto záměru však bude nepochybně velmi pracná a náročná.

Korpus DIAKORP dosáhl v roce 2000 celkového objemu 1 750 000 textových slov. Jeho součástí je mimo jiné banka transliterovaných textů (o celkovém rozsahu přibližně 100 000 textových slov) a jazyková databáze (v níž se shromažďují překlady starších českých slov, vysvětlivky jednotlivých obratů, míst v textech apod., obsažené v jednotlivých edicích).

Mluvený synchronní korpus

Mluvený synchronní korpus ORAL-PMK (*Pražský mluvený korpus*), který je samostatnou složkou ČNK, byl původně vytvářen pro účely výzkumu frekvence autentické mluvené (především obecné) češtiny a jeho rozsah a povaha byly do vysoké míry dány pragmaticky – možnostmi, které byly k dispozici. Jeho dnešní rozsah činí asi 700 000 slov. Na základě tohoto korpusu se připravuje do tisku první frekvenční slovník mluvené češtiny. Vzhledem k limitujícím faktorům je prozatím omezen na oblast Prahy a okolí (Praha ovlivňuje ostatní území nejen mediálně, ale i tím, že v ní pracují lidé z celé země). Mluvený korpus časově pokrývá období let 1988–1996 a představuje i přes svá omezení zdaleka největší a nejreprezentativnější záznam autentického mluvené češtiny. Rozsah 700 000 textových slov je podle dnešních nároků a ve srovnání s psaným synchronním korpusem malý, je však třeba mít na paměti, že odpovídá reálným možnostem. Jde především o to, že získání mluveného korpusu je mnohonásobně pracnější a dražší než vybudování korpusu psaného (viz 1.2). Na základě pravidel, podle kterých byl sestaven PMK, byl vytvořen BMK (*Brněnský mluvený korpus*) s asi 500 0000 tvary, který umožní srovnávat jazykovou situaci těchto dvou velkých měst. V současnosti probíhá sběr mluvených textů v Čechách, na Moravě a ve Slezsku, v budoucnu se plánují další velké sběry. Tyto korpusy umožní popsat jazykovou situaci mluveného jazyka, která je dodnes víceméně jen odhadována.

Projekt mluveného korpusu (PMK a BMK) sledoval reprezentativní zastoupení čtyř hlavních sociolingvistických proměnných: pohlaví mluvčích (muž–žena), věku (vyšší–nižší s hranicí kolem 35 let, ale s vyloučením jazyka nedospělé mládeže), vzdělání (nižší–vyšší) a typu textu, resp. nahrávky (formální, tj. podle předem připravených širokých otázek, nebo neformální, tj. neřízený dialog dvou mluvčích, kteří se znají). Tak vzniklo několik desítek kombinací těchto čtyř indexů, které se nahrávkami naplňovaly tak, aby se dosáhlo vyvážené, proporcionální podoby. Nahrávky se pak manuálně přepisovaly standardním způsobem do počítače a u PMK anotovaly.

2.3 Reprezentativnost ČNK

Vymezení textových zdrojů korpusu a kvantitativní pohled na jejich strukturu jsou spojené nádoby. Každá kvantitativně vyjádřitelná část korpusu pochází z určitého konkrétního okruhu zdrojů a naopak každý okruh zdrojů má určité oprávnění být v korpusu zastoupen určitým kvantem textů. Při seriózním přístupu ke stavbě korpusu proto nelze volit ani okruhy zdrojů, ani míru jejich zastoupení pouze intuitivně a tím spíše ne náhodně.

Ideální textový korpus zahrnuje z hlediska matematické statistiky neuskutečnitelné soustředění naprosto všech textů, které bychom chtěli mít k dispozici. Každý reálný korpus – přes svou rozsáhlost – tak zůstává tím, co statistikové nazývají vzorek, výběr. Úvaha o reprezentativnosti vzorku je pak úvahou o míře jistoty (pravděpodobnosti), že daný vzorek, výběr, zahrne také ty či ony prvky nebo jevy. Je obecně známo, že reprezentativnost neroste lineárně s pouhým zvětšováním rozsahu, ale že vždy závisí především na tom, co sledujeme, na pravděpodobnostních charakteristikách (rozloženích) výskytů sledovaných prvků nebo jevů. V případě korpusu jde však o prvky a jevy v takovém množství, že statistické (pravděpodobnostní) konstruování struktury zdrojů – i rozsahu – korpusu tak, aby byl naprosto všestranně reprezentativní, je vyloučeno.

Reprezentativnost korpusu je však možno chápat také z hlediska jeho možných využití. Nelze-li strukturu stavby všestranně reprezentativního korpusu odvozovat od statistických charakteristik jednotlivých – pro korpus vnitřních – jazykových prvků či jevů, je třeba se pokusit o konstrukci struktury textového korpusu z hledisek vnějších. Tímto směrem vedly i hlavní úvahy o koncepci reprezentativnosti nejrozsáhlejší složky ČNK – stomilionového synchronního korpusu psaných textů SYN2000.

Vzhledem k tomu, že neexistuje univerzální, všeobecně uznávaná představa o využití korpusů (neboť neexistuje ani představa o všech možných budoucích požadavcích kladených na korpusy jako takové), staly se východiskem pro ČNK možná očekávání ze strany dnes uvažovaných potenciálních zájemců o práci s korpusem. Při hledání optimální struktury zdrojů (současných psaných textů) pro ČNK jsme se opírali o průzkumy, jejichž dat bylo možno po přetřídění a dalších propočtech využít jednak ke konfrontaci s původním intuitivním (pracovním) návrhem, jednak k číselnému zakotvení návrhu nového.

Na základě dosavadních průzkumů bylo dosaženo několika stupňů poznání. Pro současné členění textových zdrojů ČNK se stal výchozím bodem první úrovně výsledek průzkumu o poměru *čtení knih, časopisů a novin* českou populací (Opinion Window Prague 1996). Odtud vyplynulo zastoupení novin (denního tisku) v SYN2000 v rozsahu 60 %.

Pro druhou úroveň členění zdrojů ČNK (členění nenovinových textů na imaginativní a informativní) byl přijat vyvážený souhrn výsledků tří různých průzkumů týkajících se porovnání zájmu o naučnou a krásnou literaturu ze strany různých skupin čtenářů. Šlo o průzkum struktury *výpůjček ve veřejných knihovnách* (Statistika veřejných knihoven ČR, 1996), o průzkum struktury *zájmu o žánry nových knižních edic* ze strany čtenářů nových knih (Halada a Jeřábek, 1994) a o průzkum *vztahu mezi fondy a výpůjčkami ve veřejných knihovnách* (Struktura fondů a výpůjček v knihovnách, 1994). Z výsledků bylo odvozeno procentuální zastoupení naučné literatury (informativních nenovinových textů) – 25 %, a krásné literatury (imaginativních textů) – 15 %.

Třetí, detailní úroveň členění zdrojů ČNK (vnitřní členění informativních a imaginativních textů) se opírala o výsledky průzkumu *struktury zájmu o naučnou literaturu* (Výpůjčky literatury podle tematického zaměření – Statistika veřejných knihoven ČR 1996), průzkumu *struktury zájmu o literaturu* ze strany čtenářů nových knih (Halada a Jeřábek, 1994), a průzkumu *struktury katalogu domácích periodik* (Šulc, 1999).

Vypočtená procentuální zastoupení jednotlivých oborových oblastí a jejich další členění byla ještě korigována z věcných hledisek, a to vzhledem k (1) požadavku co nejširšího záběru ČNK, (2) omezené dostupnosti mluveného (nejen obecného jazyka) a (3) univerzálnosti využití ČNK. Výslednou strukturu, užitou v korpusu SYN2000, shrnuje v hlavních rysech následující tabulka.

Tab. č. 1 *Struktura textů v korpusu SYN2000*

TYP TEXTU	Podíl [%]
IMAGINATIVNÍ TEXTY	15,00
krásná literatura	12,04
poezie	0,81
drama	0,21
próza	11,02
jiné imaginativní texty	0,36
přechodové pásmo	2,60
INFORMATIVNÍ TEXTY	85,00
publicistika	60,00
odborné texty	25,00
vědy o umění	3,48
sociální vědy	3,67
právo a bezpečnost	0,82
přírodní vědy	3,37
technika	4,61
ekonomie a řízení	2,27
víra, náboženství	0,74
životní styl	5,55

administrativa	0,49

Tato struktura se odráží i v třídění textů do jednotlivých oborů, k němuž dochází při jejich zařazování do ČNK, a v jejich anotaci. Systém třídění a anotace užívaný v ČNK se v základních rysech opírá o standardy vypracované skupinou TEI a skupinou EAGLES, jejichž doporučení jsou podporována Evropskou unií. Podobně jako většina korpusových projektů však i ČNK používá pro třídění a anotaci korpusového materiálu vlastní, v současné době už ustálenou sadu charakteristik a vlastní škálování uvnitř těchto charakteristik. Tato sada, implementovaná v korpusu SYN2000, je založena na původním návrhu Františka Čermáka, modifikovaném podle Deweyho desetinného třídění a podle zkušeností zahraničních korpusových projektů; celkově se v ní pracuje s 11 kategoriemi a 112 rysy uvnitř těchto kategorií.

2.4 Způsob budování ČNK

2.4.1 Získávání textů

Texty pro ČNK jsou získávány celkem pěti různými způsoby:

1. prostřednictvím smluv s nakladateli a vydavateli
2. využíváním textů dostupných na internetu
3. skenováním
4. manuálním přepisem
5. darem od autorů

Naprostá většina textů, které dnes tvoří jednotlivé korpusy ČNK, byla získána v elektronické formě přímo od nakladatelů a vydavatelů. ÚČNK během své dosavadní existence uzavřel s těmito poskytovateli textů více než 300 smluv, na jejichž základě jsou poskytovány i celé ročníky periodik, jako jsou *Lidové noviny* nebo *Mladá fronta Dnes*, a produkce nakladatelství (např. Nakladatelství Lidové noviny, Mladá fronta, Atlantis, TORST aj.). Smlouvy zavazují ÚČNK k užití textů pouze pro budování ČNK a pro jejich nekomerční využívání (rozsah citací z jednotlivých textů je přitom omezen horní hranicí 100 slov).

Texty od nakladatelů a vydavatelů, pro účely ČNK dále konvertované a zpracovávané do jednotné podoby (viz níže), tvoří více než 90 % stomilionového korpusu SYN2000 a prakticky zcela uspokojují potřeby ČNK v oblasti současných centrálních i regionálních publicistických textů. Ty části synchronního korpusu, které tento relativně nejméně pracný a nákladný způsob získávání textů pokrývá jen zčásti (krásná a odborná literatura) nebo zcela okrajově (texty z oblasti různých řemesel, domácího hospodářství, běžné administrativní texty, soukromé dopisy, oznámení, letáky, návody apod.), jsou doplňovány dalšími výše uvedenými způsoby získávání textů, především skenováním (podílejícím se na celkovém rozsahu korpusu SYN2000 přibližně 7 miliony slovních tvarů, tj. zhruba 7 %) a přepisem (podíl v rámci SYN2000 je zhruba 330 000 slovních tvarů, tj. 0,3 %). Textů darovaných přímo autory je zanedbatelné množství (jakkoli mohou některé z nich být důležité z hlediska zastoupení jednotlivých textových typů v korpusu), nicméně texty přístupné prostřednictvím

internetu nabývají na stále větší důležitosti. Skenování a přepisování textů probíhá vzhledem ke své poměrně značné pracnosti a časové i finanční náročnosti plánovitě podle programů cíleně zaplňujících mezery ve skladbě synchronního psaného korpusu. Nejrozsáhlejší z těchto programů, zaměřený na skenování krásné literatury a vypracovaný po konzultacích s Ústavem pro českou literaturu AV ČR, představuje soubor nejčtenějších literárních děl 2. poloviny 20. století doplněný nejčtenějšími literární překlady a výběrem stěžejních dramatických děl. Součástí tohoto programu je i kompletace děl jednotlivých významných autorů (v současné době je v elektronické podobě kompletováno dílo Karla Čapka a Bohumila Hrabala).

2.4.2 Zpracování textů

Každý text vstupující do ČNK je zaevidován a ve své původní podobě (tj. v té, v jaké byl získán od nakladatele, naskenován, přepsán apod.) je uložen do textového **archivu**. Pro potřeby ČNK je třeba dále všechny získané texty konvertovat do jednotného formátu SGML, anotovat je a označkovat; takto připravené texty jsou uloženy do textové **banky** a dále se s nimi pracuje jako se *soubory*. Soubor většinou odpovídá jedné knize nebo jednomu číslu novin, ale podle potřeby může mít i jiný rozsah (např. různorodý text jedné knihy může být rozdělen do několika souborů nebo naopak celý ročník novin lze spojit do jednoho souboru). Při konverzi se v co největší míře zachovává autenticita textu (zachovává se jeho původní členění, neopravují se chyby ani zjevné překlepy apod.); jedinou výjimku tvoří odstraňování odstavců obsahujících cizojazyčný text, tabulky, vzorce apod.

Formát SGML, v němž jednotlivé soubory vstupují do banky, je přizpůsoben potřebám ČNK: používá vlastní DTD (document type definition – popis SGML dokumentu) a jednotné kódování češtiny (ISOLatin2, tj. ISO-8859-2). Každý soubor je jednoznačně identifikován svým *jménem* (je zajištěno, že žádné dva soubory v bance nemohou být stejně pojmenovány) a opatřen *hlavičkou*, v níž jsou uvedeny všechny relevantní technické informace o konverzi (mimo jiné i o tom, kdo a kdy prováděl jednotlivé části konverze). Hlavička souboru je pro běžného uživatele skrytá, technickým pracovníkům však pomáhá dohledávat a opravovat eventuální chyby.

Každý soubor v bance je anotován a strukturován do dokumentů, odstavců, vět a slov. V rámci souboru jsou dokumenty, odstavce a věty číslovány, což spolu s jednoznačným pojmenováním souborů umožňuje jednoznačnou identifikaci všech objektů v bance až na úrovně jednotlivých vět.

Ukázka souboru ve formátu SGML

Soubor začíná hlavičkou, ve které je zaznamenán postup konverzí, dále následuje hlavička dokumentu s označením typu textu a konečně tělo dokumentu s vlastním textem, rozděleným do odstavců a vět.

```
<!DOCTYPE csts PUBLIC "-//ICNC//DTD CSTS ver. 09//EN">
<csts lang=cs>
<h>                                  <!-- začátek hlavičky souboru -->
<source>Skenování                    <!-- zdroj, způsob získání textu -->
</source>
<markup>
<mauth>Michal Křen                   <!-- autor konverze, -->
<mdate>2000-02-07                    <!-- její datum -->
<mdesc>foreign languages cleanup     <!-- a typ -->
</markup>
<markup>
<mauth>Jarka Hlaváčová               <!-- autor konverze, -->
<mdate>07.02.2000                    <!-- její datum -->
<mdesc>1st -> 2nd intermediate format <!-- a typ -->
</markup>
<markup>
<mauth>Michal Křen                   <!-- autor konverze, -->
<mdate>07-Feb-2000                   <!-- její datum -->
<mdesc>csts style markup             <!-- a typ -->
</markup>
</h>                                 <!-- konec hlavičky souboru -->
<a>                                  <!-- začátek hlavičky dokumentu -->
<mod>S                               <!-- korpus (synchronní) -->
<txtype>COL                          <!-- typ textu (soubor povídek) -->
<genre>JUN                           <!-- žánr (literatura pro děti a mládež)-->
<verse>NRH                           <!-- veršovanost (neveršovaný) -->
<med>B                               <!-- médium (kniha) -->
<authsex>M                           <!-- pohlaví autora -->
<lang>CZE                            <!-- jazyk -->
<transsex>NA                         <!-- pohlaví překladatele -->
<srclang>NO                          <!-- původní jazyk -->
<temp>1978                           <!-- rok vydání -->
<firsted>1960                        <!-- rok prvního vydání -->
<authname>Werich, Jan                <!-- jméno autora -->
<transname>NA                        <!-- jméno překladatele -->
<opus>fimfarum                       <!-- jednoznačná identifikace souboru -->
<id>001                              <!-- číslo dokumentu v rámci souboru -->
</a>                                 <!-- konec hlavičky dokumentu -->
<c>                                  <!-- začátek těla dokumentu -->
<p n=1>                              <!-- označení odstavce s jeho číslem v rámci dokumentu -->
<s id="S/B/1978/fimfarum:001-p1s1">  <!-- označení věty s jejím číslem v rámci odstavce -->
<f upper>KRÁLOVNA                    <!-- slovo psané kapitálkami -->
<f upper>KOLOBĚŽKA
<f upper>PRVNÍ
<p n=2>                              <!-- začátek druhého odstavce -->
<s id="S/B/1978/fimfarum:001-p2s1">  <!-- začátek další věty -->
<f cap>V                             <!-- slovo začínající velkým písmenem -->
<f>jednom                            <!-- „obyčejné" slovo -->
<f>malilinkém
<f>království
<f>žili
<f>mlynář
<f>a
<f>rybář
<D>                                  <!-- následující znak nebyl oddělen mezerou -->
<d>.                                 <!-- interpunkce -->
<s id="S/B/1978/fimfarum:001-p2s2">  <!-- začátek druhé věty druhého odstavce-->
<f cap>Bydlili
<f>blízko
<f>sebe
<D>
<d>,
<f>pochopitelně
<f>u
<f>řeky
<D>
<d>.
```

Jak je zřejmé z uvedeného schématu, každý soubor se skládá z jednoho nebo více *dokumentů.* Každý dokument je opatřen hlavičkou, podobně jako soubor, avšak v hlavičce dokumentu je uchovávána *anotace* (informace o autorovi, typu textu, žánru, roku vydání apod.). Dokumenty tvoří menší logické celky, než je celý soubor (soubor v bance může být tvořen např. knihou povídek, z nichž každá je samostatným dokumentem, nebo jedním číslem novin, v němž dokumentům odpovídají jednotlivé články). Členění souborů na dokumenty je do značné míry volné a v rámci technických možností závisí na uvážení lingvisty, který text pro ČNK zpracovává, a na konkrétních vlastnostech souboru (ve vstupním textu například není vždy možno jednoznačně rozpoznat hranice jednotlivých článků nebo povídek). Velikost souboru ani dokumentu není nijak omezena. Rozčlenění textu na dokumenty je součástí konverze z tzv. 1. do 2. meziformátu. Jemnost vnitřního členění větších textů (např. novinových) však závisí na pracovních a časových možnostech týmu ČNK.

Členění dokumentů na odstavce je dáno již původním textem a při konverzi je věnována soustavná pozornost tomu, aby bylo zachováno (výjimkou je výše zmíněné převážně automatické odstraňování odstavců obsahujících cizojazyčný text, tabulky apod.). Členění odstavců na jednotlivé *věty* a *slova* se provádí plně automaticky při převodu z 2. meziformátu do SGML.

Konverze textových souborů probíhá ve třech hlavních fázích:

1. převod textu z původní podoby do 1. meziformátu,
2. převod textu 1. meziformátu do 2. meziformátu,
3. převod textu z 2. meziformátu do formátu SGML.

Jako text v 1. meziformátu se v rámci ČNK označuje ASCII text v kódování CP 1250 (kódová stránka Windows), který může navíc obsahovat některé speciální značky. Pomocí těchto značek jsou kódovány informace, které se daly zjistit z původního textu, ale které by se převodem do prostého ASCII textu ztratily. Jde zejména o typografické značky (vyznačující např. tučné a podtržené písmo, kurzívu, nadpisy, horní a dolní index apod.) a o nečeské znaky s diakritickými znaménky, které nejsou zahrnuty v kódování CP 1250. Převod textu do 1. meziformátu je poměrně komplikovaný a velmi při něm záleží na formátu vstupních dat, konkrétně na editoru, v němž byly původní texty vytvořeny. Pro formáty HTML, RTF, T602 a WordPerfect byly v ČNK vyvinuty spolehlivé konverzní programy, které generují přímo 1. meziformát, a to i se speciálním kódováním nečeských znaků a typografických značek; i v těchto případech je však vždy nutné výsledek zkontrolovat a zpravidla ještě napsat pro každý převod speciální program, který odstraní některé nevhodné jevy, jako např. oddělování částí textu pomocí řádek složených z pomlček nebo podtržítek, označování stran a obrázků apod.

Je-li vstupní formát natolik složitý, že by bylo obtížné psát vlastní konverzní program (např. editor MS Word nebo programy DTP), nezbývá než text pomocí těchto programů načíst a převést do některého z jednodušších, snáze konvertovatelných formátů. I tento postup lze do značné míry automatizovat, především pomocí maker MS Wordu nebo užitím programu WinBatch (v případě některých formátů DTP). Problémy s některými vstupními formáty však přesto trvají: ne vždy lze totiž použít některý z výše uvedených způsobů a občas je tedy nutné převádět texty ručně. Vzhledem k pracnosti ručního převodu se této možnosti využívá pouze

v případech, kdy je text z nějakého důvodu vzácný a nelze se bez něho obejít ani jej získat jinak.

Hlavním důvodem převodu textu do 1. meziformátu je především to, aby všechny texty získaly jednotnou podobu a veškeré další zpracování mohlo probíhat jednotně, bez ohledu na původní formát textů. Výsledek tohoto dalšího zpracování (2. meziformát) se od 1. meziformátu na první pohled výrazně neliší: jde rovněž o ASCII text, navíc však obsahující hlavičku SGML s lingvistickou anotací textu. Anotaci provádějí většinou lingvisté na PC v prostředí MS FoxPro (to je také důvod, proč je 1. meziformát v kódování CP 1250) a zapisují ji přitom do databáze (podrobně o evidenci textů v databázi viz 2.4.3). Druhý meziformát pak vzniká exportováním anotace z databáze do hlaviček textů a jejím přidáním k textům. U velkých kvant textových souborů s jednotnou anotací (například noviny) lze anotování automatizovat; v takovém případě se anotace do databáze zapíše dávkově, hlavičky se k textům generují automaticky a převod do 2. meziformátu se provede přímo. Důležitou součástí převodu textů z 1. do 2. meziformátu je také výše zmíněné členění nebo naopak spojování textů do logických celků (*textů* a *dokumentů*) a případné ruční odstraňování některých částí textu, které by pozdější automatická detekce neodhalila (např. tabulky, nakladatelské údaje (tiráž), seznamy dodatečně připojené k textu, které netvoří jeho organickou součást apod.).

Posledním krokem konverze je převod z 2. meziformátu do formátu SGML. Tento proces je již plně automatizován, texty při něm procházejí nejprve *tokenizerem* (programem pro segmentaci textu do vět a slov), který je převede do SGML, a poté ještě několika *čisticími* a *kontrolními programy*. K těmto programům patří zejména program na detekci cizích jazyků, který pracuje s již hotovými texty a odstraňuje z nich odstavce psané převážně cizím jazykem. Je třeba zdůraznit, že program pracuje pouze s celými odstavci (tj. celý odstavec buď v textu ponechá, nebo ho vymaže), a nedochází tedy k mazání krátkých cizojazyčných citací, které jsou součástí okolního českého textu.

Z hotových textů se dále odstraňují různé části netextové povahy, jako např. číselné tabulky nebo obrázky. Detekce těchto součástí textu je poměrně složitá, mimo jiné i proto, že dosud neexistuje plný konsensus v tom, které netextové části je vhodné odstranit a které nikoliv; program proto odstraňuje zásadně pouze ty části, jejichž lingvistická hodnota je zcela zanedbatelná (např. burzovní tabulky).

Posledním čisticím programem je program na odstraňování duplicitních textů, které se (nejčastěji v podobě několika totožných nebo jen nepatrně odlišných verzí téhož článku) poměrně často vyskytují ve zdrojových textech z vydavatelství novin a časopisů, což jde zřejmě na vrub textů pocházejících ze společného agenturního zdroje, které nejsou novináři dále upravovány. Program prochází texty a v rozsahu několika čísel téhož časopisu nebo novin vyhledává na úrovni dokumentů duplicitní články a odstraňuje je. I problém duplicity je poměrně komplikovaný, a je proto i v tomto případě nastaven tak, aby vymazával jen ty dokumenty, které se plně nebo téměř plně shodují.

Posledním krokem převodu z 2. meziformátu do SGML je značkování (tagování) textů, tj. přidávání morfologických značek a lemmat k jednotlivým slovním tvarům v textech (podrobněji viz 2.4.5).

2.4.3 Evidence textů

Všechny texty, které jsou ukládány do archivu ČNK, je třeba evidovat a průběžně zaznamenávat stav jejich zpracování. K tomu slouží databáze *Evidence* vytvořená a udržovaná ve vývojovém prostředí Visual FoxPro, kterou tvoří několik navzájem propojených tabulek. Do výchozí tabulky se zapisují údaje o všech textech, které se podaří jakýmkoliv způsobem získat; v zásadě jde o údaje trojího typu: 1. bibliografická data (název díla, jméno autora, nakladatel, rok vydání atd.), 2. formální charakteristika textového souboru (editor, v němž je text napsán, kódová stránka) a 3. administrativní údaje o textu (jméno souboru nebo souborů, v nichž je příslušný text uložen; číslo disku, na kterém jsou soubory zapsány; jméno pracovníka, který převzal text k anotaci; místo, kde je uložen meziformát, stupeň rozpracování textu apod.).

Podle výchozí tabulky je průběžně aktualizována tabulka organizací poskytujících texty ČNK; podle ní si také programátoři vybírají soubory ke zpracování. Po převedení do 1. meziformátu jsou tyto soubory předávány lingvistům k vnější lingvistické anotaci, v jejímž rámci je text zařazen do kategorií podle několika hledisek a je mu přiřazen kód, který ho jednoznačně identifikuje v rámci banky. Anotaci lingvisté provádějí ve svých osobních tabulkách, kam si zkopírují příslušné záznamy z výchozí tabulky, do níž se přitom automaticky zapíše jejich jméno a datum převzetí záznamů, čímž se zabrání dvojí anotaci týchž textů. Obslužný program, pomocí něhož lingvisté zanášejí jednotlivé údaje do osobních tabulek, je opatřen řadou automatických kontrol vylučujících základní chyby při zápisu nových údajů a obsahuje i postupy, které umožňují anotovat více záznamů najednou stejnými značkami, ulehčují práci s kopírováním údajů aj. Během anotace se podle aktuálního stavu značení automaticky vyplňuje speciální položka *Stav* v databázi Evidence.

Po dokončení anotace jsou všechny správně vyplněné záznamy převedeny do výsledné tabulky databáze Evidence, přičemž se automaticky kontroluje, zda kódy nově přidávaných souborů nebyly použity již dříve a zda je tedy není nutno změnit. Současně se do výchozí tabulky zapíše, že příslušný záznam je už ve výsledné tabulce a jeho zpracování je dokončeno. Záznamy o zpracovaných souborech se zkopírují do záložní tabulky obsahující záznamy o všech souborech, které daný lingvista úspěšně označil a převedl do výsledné tabulky, a současně se tyto záznamy vymažou z jeho osobní tabulky.

Tabulka Evidence obsahuje všechny dostupné údaje o každém textu, který je v bance ČNK. Její struktura vypadá takto:

Tab. č. 2 Struktura tabulky Evidence

Název pole	**Popis**
DISK_CISLO	číslo disku, na kterém je soubor s textem uložen
SOUB_NAZEV	název souboru
EDITOR	editor, ve kterém je napsána původní podoba textu
KOD	původní kódová stránka
OPUS	kód jednoznačně určující text
MEZIFORMAT	číslo CD, kde je uložena záloha meziformátu
AUTORI	autoři textu
JEDNOTKA	název textu

ČÁST	případný údaj o části titulu, ve které je text obsažen
TITUL	titul, ve kterém je text obsažen
PODTITUL	podtitul
AUTOR_NAD	autor titulu, ve kterém je text obsažen
ISBN_ISSN	ISBN nebo ISSN
NAKLADATEL	jméno a adresa nakladatele
ORGANIZACE	jméno a adresa organizace, která poskytla text ČNK
VYDANI	pořadí vydání
MÍSTO_VYD	místo vydání
PREKLADATE	překladatel
BIBL_POZN	bibliografická poznámka
MOD	modus (synchronní – diachronní – paralelní)
TXTYPE	typ textu
GENRE	žánr
VERSE	informace o tom, zda text má veršovou formu
MED	medium, na kterém byl text vydán
AUTHSEX	pohlaví autora
LANG	jazyk
TRANSSEX	pohlaví překladatele
SRCLANG	jazyk původního textu (pouze u překladů)
TEMP	rok vydání
FIRSTED	rok 1. vydání
STAV	stupeň rozpracovanosti textu
POZNAMKA	jakákoli poznámka, např. o problémech při zpracování textu
KOREKCE	informace, zda text prošel v nakladatelství korekturami

2.4.4 Značkování

Značkování (tagování) je proces, při němž jsou texty vstupující do korpusu opatřovány (anotovány) doprovodnými informacemi, které jsou formálně vyjádřeny prostřednictvím *značek* (*tags*). Tyto značky jsou trojího druhu: administrativní či správní (vnější anotace), strukturní a lingvistické (vnitřní anotace).

Administrativní (správní) značky zachycují administrativní údaje o každém textu soustředěné v tzv. *hlavičce* a obsahují zejména informace o původu, autorství, typu a zdroji textu. Níže je uveden příklad takové hlavičky:

<csts lang=cz>
<h>
<source>VES 1992
</source>
<markup>
<mauth>Jan Hajic
<mdate>Sun Mar 5 00:51:04 2000
<mdesc>Morphology; parameters: RootOut=0, EndOut=0, AllTags=0,LemmaTagUpdate=0, ForceLemmaUpdate=0
<mdesc>Morphology; parameters, set 2: MDCopy=0
<mdesc>MA: syn/sem/sty: :_W_T_B/;_G_Y_S_E_R_K_H_U_L/,_x_s_a_n_h_e_l_v_t/

<mdesc>MA: output: desc: ^Yes, la:/lc:_s_a_n_h_e_l_v, va:/vc:-3-4-5-6-7-9;
</markup>
<markup>
<mauth>Michal Kren
<mdate>1997-02-17
<mdesc>duplicate documents cleanup
</markup>
<markup>
<mauth>Michal Kren
<mdate>1997-02-13
<mdesc>foreign languages cleanup
</markup>
<markup>
<mauth>Martin Tengler
<mdate>1996
<mdesc>conversion to csts DTD v 0.9 (0023)
</markup>
<markup>
<mauth>Lubos Ures
<mdate>28-Nov-1995
<mdesc>csts style markup
</markup>
<markup>
<mauth>Jan Holub
<mdate>1995
<mdesc>raw text -> intermediate clean text
</markup>
<markup>
<mauth>Petr Kubicek
<mdate>28-Apr-1998
<mdesc>modified csts format v. 0.9
</markup>
</h>
<doc id=001 file="S/J/1992/vesm9211">
<a>
<mod>S
<txtype>SCI
<genre>NAT
<verse>NRH
<med>J
<authsex>Y
<lang>CZE
<temp>1992
<firsted>1992
<authname>Y
<opus>vesm9211
<id>001

Toto záhlaví popisuje administrativní údaje o textech z 11. čísla časopisu *Vesmír* z roku 1992, především různé typy zpracování textu, data zpracování, jména pracovníků, kteří texty zpracovávali, a identifikaci textu (úsek začínající značkou <doc...>).

Vlastní text (např. kniha) bývá hierarchicky členěn *strukturními značkami* např. na kapitoly, jež jsou tvořeny odstavci, přičemž každý odstavec je tvořen posloupností vět, z nichž každá je z formálního hlediska posloupností tzv. textových slov (*tokenů*), tj. výskytů slovních tvarů (slovních exemplářů), čísel, zkratek, speciálních znaků (např. znak *$* pro americký dolar) a interpunkčních znamének. Jednotlivé strukturní značky vymezují identitu daného úseku textu, tj. explicitně vyznačují, kde daný úsek textu začíná a kde končí. Úsek je uvozen tzv. otvírací značkou, která má podobu <znacka>; pak následuje příslušný úsek textu, který bývá (avšak ne nutně) posléze zakončen tzv. uzavírací (ukončovací) značkou tvaru </znacka>, nebo bezprostředně následuje další element daného typu (např. věta). Tak např. 3. odstavec vybraného textu z časopisu *Vesmír* z roku 1992 tvořený posloupností vět je strukturně značkován takto:

<p n=3> ... 3. odstavec...
<s id="S/J/1992/vesm9211:001-p3s1">
... 1. věta...
<s id="S/J/1992/vesm9211:001-p3s2">
... 2. věta...

...

<s id="S/J/1992/vesm9211:001-p3s6">
... předposlední věta...
<s id="S/J/1992/vesm9211:001-p3s7">
... poslední věta...

<p n=4>

Jak je z příkladu zřejmé, každý odstavec a každá věta jsou jednoznačně identifikovány.

Protože ČNK slouží především jazykovědnému výzkumu, je žádoucí, aby obsahoval více než jen údaje o struktuře jednotlivých textů, segmentovaných až na úroveň jednotlivých slov, i když samozřejmě již samotné slovní tvary, jejich konfigurace a jejich lexikální kolokace (kombinovatelnost) jsou cenným zdrojem informací. Jazykové bohatství, jež korpus poskytuje, lze však vytěžit zejména tím, že se jednotlivým slovním tvarům (dále jen slova) v textech korpusu přiřadí různé lingvistické informace. Obohacení slov v korpusových textech zejména o lingvistické údaje znamená, že v korpusu lze vyhledávat nejen podle samotných slovních tvarů a jejich konfigurací, ale i podle jejich lingvistických charakteristik; v současné podobě v korpusu SYN2000 jde o charakteristiky morfologické, včetně slovnědruhové příslušnosti. Tyto charakteristiky jsou v podobě *lingvistických značek* přiřazovány k jednotlivým slovům tzv. *morfologickým analyzátorem* – programovým modulem opatřujícím automaticky každé slovo v textu lingvistickou informací dvojího typu:

1. *Lemmatizací* je danému slovu přiřazena informace o jeho základním, slovníkovém tvaru zvaném *lemma*, popř. o více možných základních tvarech. Více základních tvarů se uvádí: (a) u forem slovnědruhově víceznačných (např. slovo *spíš*, které je buď tvarem 2. osoby jednotného čísla od slovesa *spát*, a patří tedy k lemmatu *spát*, nebo je příslovcem s lemmatem

spíš); (b) u forem, které odpovídají více základním tvarům v rámci jednoho slovního druhu a jež jsou tedy lexikálně víceznačné (např. u sloves *cenit*[1] – „ukazovat zuby„, a *cenit*[2] – „určovat cenu„). Lemmata jsou v současné verzi značkování pojata poněkud šířeji, než je obvyklé, a to zejména v tom smyslu, že některé lexikální jednotky jsou slučovány do jednoho lemmatu: např. přechýlené podoby podstatných jmen mají lemma totožné se základovým podstatným jménem, od něhož jsou odvozeny; záporné podoby sloves (např. *nekradu*) jsou zahrnuty pod lemma bez záporky (tj. *krást*) apod.

2. Vedle lemmat přiřazuje morfologický analyzátor každé formě *všechny* její *potenciální morfologické interpretace*, tj. informace o její slovnědruhové příslušnosti a morfologických vlastnostech (např. o rodu, čísle a pádu podstatných a přídavných jmen, zájmen a číslovek, o stupni přídavných jmen a příslovcí, o osobě, čísle, slovesném a jmenném rodu slovesných tvarů atd.). Morfologická interpretace daného slova je formálně vyjádřena *morfologickou značkou* tvořenou maximálně 15 údaji, z nichž každý je reprezentován jedním znakem na dané pozici, přičemž význam jednotlivých pozic je jednoznačně stanoven (např. 1. pozice obsahuje identifikaci *slovního druhu*, 2. pozice obsahuje *jemnější kategorizaci slovního druhu* vymezeného základní slovnědruhovou hodnotou na 1. pozici, 3. pozice identifikuje *jmenný rod*, 4. pozice *číslo*, 5. pozice *pád* atd. Každá morfologická značka je tak tvořena patnáctičlennou řadou znaků; např. morfologická značka

NNMS1-----A----

má tuto interpretaci: N = substantivum, N = obecné substantivum, M = maskulinum, S = singulár, 1 = nominativ, A = kladná, nikoli negativní podoba slova (protiklad typu *víra/nevíra*). Je-li dané slovo morfologicky, případně lexikálně či slovnědruhově víceznačné (např. výše zmíněný tvar *spíš* nebo tvar *prosí*, jenž může být slovesným tvarem 3. osoby jednotného i množného čísla od slovesa *prosit*), opatří je morfologický analyzátor tolika patnáctimístnými značkami, kolik má toto slovo lexikálních, slovnědruhových a morfologických významů, a to včetně příslušných lemmat (např. uvedený tvar *prosí* bude morfologicky charakterizován dvěma značkami, které se budou shodovat v hodnotách na všech pozicích s výjimkou hodnoty na pozici čísla).

Jako příklad uveďme morfologické označkování 1. věty 3. odstavce z již zmíněného textu z časopisu *Vesmír*. Neoznačkovaná věta vypadá takto:

Sir John Eccles, neurofyziolog a laureát Nobelovy ceny, a jeho přítel sir Karl Popper, filozof vědy, postupně a nelehce stvořili jednu z nejelegantnějších teorií vysvětlujících vztahy lidského mozku k duši, i jedince ke kulturnímu dědictví a opačně.

Morfologicky a slovnědruhově označkovaná věta má tuto podobu:

<p n=3>
<s id="S/J/1992/vesm9211:001-p3s1">
<f cap>Sir<MDl>sir<MDt>NNMS1-----A----<MMl>Sira<MMt>NNFP2-----A----
<MMl>sir<MMt>NNMS1-----A----
<f cap>John<MDl>John<MDt>NNMS1-----A----<MMl>John<MMt>NNMS1-----A----
<f cap>Eccles<MDl>Eccles<MDt>NNMS1-----A----<MMl>Eccles<MMt>NNMS1-----A----
<D>
<d>,<MDl>,<MDt>Z:----------<MMl>,<MMt>Z:----------

<f>neurofyziolog<MDl>neurofyziolog<MDt>NNMS1-----A----
<MMl>neurofyziolog<MMt>NNMS1-----A----
<D>
<d>,<MDl>,<MDt>Z:----------<MMl>,<MMt>Z:----------
<f>laureát<MDl>laureát<MDt>NNMS1-----A----<MMl>laureát<MMt>NNMS1-----A----
<f cap>Nobelovy<MDl>Nobelův<MDt>AUFS2M---------<MMl>Nobelův<MMt>AUFP1M-
--------<MMt>AUFP4M---------<MMt>AUFP5M---------<MMt>AUFS2M---------
<MMt>AUIP1M---------<MMt>AUIP4M---------<MMt>AUIP5M---------<MMt>AUMP4M-
----
<f>ceny<MDl>cena<MDt>NNFS2-----A----<MMl>cena<MMt>NNFP1-----A----
<MMt>NNFP4-----A----<MMt>NNFP5-----A----<MMt>NNFS2-----A----
<MMl>cena<MMt>NNFP1-----A----<MMt>NNFP4-----A----<MMt>NNFP5-----A----
<MMt>NNFS2-----A----
<D>
<d>,<MDl>,<MDt>Z:----------<MMl>,<MMt>Z:----------
<f>a<MDl>a<MDt>J^----------<MMl>a<MMt>J^----------<MMl>a<MMt>NNNXX-----
A----
<f>jeho<MDl>jeho<MDt>PSXXXZS3-------<MMl>jeho<MMt>PSXXXZS3-------
<MMl>on<MMt>PPYS2--3-------<MMt>PPYS4--3-------
<f>přítel<MDl>přítel<MDt>NNMS1-----A----<MMl>přítel<MMt>NNMS1-----A----
<f>sir<MDl>sir<MDt>NNMS1-----A----<MMl>sir<MMt>NNMS1-----A----
<f cap>Karl<MDl>Karl<MDt>XX----------<MMl>Karl<MMt>XX----------
<f cap>Popper<MDl>Popper<MDt>NNMS1-----A----<MMl>Popper<MMt>NNMS1-----A--
--
<D>
<d>,<MDl>,<MDt>Z:----------<MMl>,<MMt>Z:----------
<f>filozof<MDl>filozof<MDt>NNMS1-----A----<MMl>filozof<MMt>NNMS1-----A----
<f>vědy<MDl>věda<MDt>NNFS2-----A----<MMl>věda<MMt>NNFP1-----A----
<MMt>NNFP4-----A----<MMt>NNFP5-----A----<MMt>NNFS2-----A----
<D>
<d>,<MDl>,<MDt>Z:----------<MMl>,<MMt>Z:----------
<f>postupně<MDl>postupně<MDt>Dg-------1A----<MMl>postupně<MMt>Dg-------1A----
<f>a<MDl>a<MDt>J^----------<MMl>a<MMt>J^----------<MMl>a<MMt>NNNXX-----
A----
<f>nelehce<MDl>lehce<MDt>Dg-------1N----<MMl>lehce<MMt>Dg-------1N----
<f>stvořili<MDl>stvořit<MDt>VpMP---XR-AA---<MMl>stvořit<MMt>VpMP---XR-AA---
<f>jednu<MDl>jeden`1<MDt>ClFS4----------<MMl>jeden`1<MMt>ClFS4----------
<f>z<MDl>z<MDt>RR--2----------<MMl>z<MMt>RR--2----------<MMl>z<MMt>NNNXX-
----A----
<f>nejelegantnějších<MDl>elegantní<MDt>AAFP2----3A----
<MMl>elegantní<MMt>AAFP2----3A----<MMt>AAFP6----3A----<MMt>AAIP2----3A----
<MMt>AAIP6----3A----<MMt>AAMP2----3A----<MMt>AAMP6----3A----<MMt>AANP2-
---3A----<MMt>AANP6----3A----
<f>teorií<MDl>teorie<MDt>NNFP2-----A----<MMl>teorie<MMt>NNFP2-----A----
<MMt>NNFS7-----A----
<f>vysvětlujících<MDl>vysvětlující<MDt>AGIP2-----A----
<MMl>vysvětlující<MMt>AGFP2-----A----<MMt>AGFP6-----A----<MMt>AGIP2-----A----

<MMt>AGIP6-----A----<MMt>AGMP2-----A----<MMt>AGMP6-----A----<MMt>AGNP2---
--A----<MMt>AGNP6-----A----
<f>vztahy<MDl>vztah<MDt>NNIP4-----A----<MMl>vztah<MMt>NNIP1-----A----
<MMt>NNIP4-----A----<MMt>NNIP5-----A----<MMt>NNIP7-----A----
<f>lidského<MDl>lidský<MDt>AAIS2----1A----<MMl>lidský<MMt>AAIS2----1A----
<MMt>AAMS2----1A----<MMt>AAMS4----1A----<MMt>AANS2----1A----
<f>mozku<MDl>mozek<MDt>NNIS2-----A----<MMl>mozek<MMt>NNIS2-----A----
<MMt>NNIS3-----A----<MMt>NNIS5-----A----<MMt>NNIS6-----A----
<f>k<MDl>k<MDt>RR--3----------<MMl>k<MMt>RR--3----------
<MMl>k<MMt>NNNXX-----A----
<f>duši<MDl>duše<MDt>NNFS6-----A----<MMl>duch<MMt>NNMP1-----A----
<MMt>NNMP5-----A----<MMl>duše<MMt>NNFS3-----A----<MMt>NNFS4-----A----
<MMt>NNFS6-----A----
<D>
<d>,<MDl>,<MDt>Z:-------------<MMl>,<MMt>Z:-------------
<f>i<MDl>i<MDt>J^-------------<MMl>i<MMt>J^-------------<MMl>i<MMt>NNNXX-----
A----
<f>jedince<MDl>jedinec<MDt>NNMS4-----A----<MMl>jedinec<MMt>NNMP4-----A----
<MMt>NNMS2-----A----<MMt>NNMS4-----A----
<f>ke<MDl>k-1<MDt>RV--3----------<MMl>k-1<MMt>RV--3----------
<f>kulturnímu<MDl>kulturní<MDt>AANS3----1A----<MMl>kulturní<MMt>AAIS3----1A-
---<MMt>AAMS3----1A----<MMt>AANS3----1A----
<f>dědictví<MDl>dědictví<MDt>NNNS3-----A----<MMl>dědictví<MMt>NNNP1-----A----
<MMt>NNNP2-----A----<MMt>NNNP4-----A----<MMt>NNNP5-----A----<MMt>NNNS1---
--A----<MMt>NNNS2-----A----<MMt>NNNS3-----A----<MMt>NNNS4-----A----
<MMt>NNNS5-----A----<MMt>NNNS6-----A----
<d>-<MDl>-<MDt>Z:-------------<MMl>-<MMt>Z:-------------
<f>a<MDl>a<MDt>J^-------------<MMl>a<MMt>J^-------------<MMl>a<MMt>NNNXX-----
A----
<f>opačně<MDl>opačně<MDt>Dg-------1A----<MMl>opačně<MMt>Dg-------1A----
<D>
<d>.<MDl>.<MDt>Z:-------------<MMl>.<MMt>Z:-------------
</s>
</p>

Jak je z uvedeného příkladu zřejmé, každý token předchází značka <f> uvozující slovo (tedy slovní tvar, resp. formu) nebo značka <d> uvozující interpunkční znaménko; značka přitom může být obohacena o atribut (např. <f cap> znamená, že příslušné slovo začíná velkým písmenem). Za touto značkou následuje slovní tvar a po něm jeho lemma uvozené značkou <MDl>. Po lemmatu následuje příslušná morfologická značka uvozená strukturní značkou <MDt>. Jak lemma, tak morfologická značka jsou vybrány z lemmat a morfologických a/nebo slovnědruhových variant, jež následují a jež jsou uvozeny strukturní značkou <MMl>, resp. <MMt>. Morfologické značkování lze názorně demonstrovat např. označkováním slovního tvaru *dědictví* z uvedené věty:

<f>dědictví<MDl>dědictví<MDt>NNNS3-----A----<MMl>dědictví<MMt>NNNP1-----A----
<MMt>NNNP2-----A----<MMt>NNNP4-----A----<MMt>NNNP5-----A----<MMt>NNNS1---
--A----<MMt>NNNS2-----A----<MMt>NNNS3-----A----<MMt>NNNS4-----A----
<MMt>NNNS5-----A----<MMt>NNNS6-----A----

Jeho lemmatem je základní tvar *dědictví* s náležitou morfologickou značkou

<MDt>NNNS3-----A----

(tedy kladné substantivum v dativu singuláru).

Ze všech teoreticky přípustných interpretací (v tomto případě vlastně pouze kombinací různých čísel a pádů) přiřazených tvaru *dědictví* morfologickým analyzátorem jsme uvedenou jedinou náležitou interpretaci schopni vybrat jen na základě konkrétního kontextu. Je tedy vidět, že opatření každého slova veškerými jeho potenciálními morfologickými interpretacemi tvoří pouze nutný předstupeň k určení náležité morfologické interpretace v daném kontextu. V konkrétním textu má totiž každé slovo téměř vždy jen jedinou morfologickou, slovnědruhovou a lexikální interpretaci, která přirozeně uživatele zajímá především. Výběr této interpretace ze všech interpretací poskytnutých morfologickým analyzátorem je cílem procedury zvané *disambiguace* (zjednoznačnění; používá se i tvar „desambiguace“).

2.4.5 Automatická morfologická analýza a disambiguace

Náležitě zaučenému a vycvičenému pracovníkovi (*anotátorovi*) nečiní „ruční“ disambiguace větší problémy, ač existují samozřejmě případy, kdy je výběr správné morfologické a slovnědruhové interpretace obtížný i pro člověka. Při obrovských objemech dat obsažených v korpusu však není v lidských silách disambiguovat texty ručně, a je tedy třeba je disambiguovat automaticky, softwarovými nástroji.

Co nejpřesnější a nejméně chybová automatická disambiguace korpusových textů je jedním z vůbec nejnáročnějších úkolů a největších výzev matematické lingvistiky, a to pro každý přirozený jazyk. Je to úkol nesrovnatelně složitější než automatická lemmatizace a automatická morfologická analýza, neboť výběr správné morfologické a slovnědruhové interpretace (formálně reprezentované příslušnou značkou) z více teoreticky možných interpretací závisí na konkrétním kontextu, v němž se dané slovo v textu vyskytuje, přičemž tu mohou hrát roli nejen faktory syntaktické, ale i sémantické. Úspěšnost automatické disambiguace je ovlivněna jednak zvolenými softwarovými nástroji, jednak morfologickou a slovnědruhovou víceznačností jazyka, v němž jsou značkované texty psány, a v neposlední řadě také kvalitou (chybovostí) samotných značkovaných textů. Čeština je jazyk s velmi vysokou mírou víceznačnosti jak morfologické, tak slovnědruhové: existuje v ní přes 1000 tříd systémové homonymie (jednu takovou typickou systémovou třídu tvoří např. nominativ, akuzativ a vokativ singuláru substantiv středního rodu skloňovaných podle vzoru *město*) a mimoto se čeština vyznačuje ve značné míře i homonymií náhodnou, která není dána vnitřním jazykovým systémem češtiny. Z toho vyplývá, že úkol automaticky morfologicky disambiguovat český text je mimořádně složitý, a to i v porovnání s ostatními slovanskými jazyky. Vzhledem ke složitosti syntaktické a sémantické struktury přirozeného jazyka obecně (a češtiny zvláště) není dosavadními prostředky možné dosáhnout automatické disambiguace korpusu bez chyb.

Obecně existují dvě základní metody automatické disambiguace textů:

1. *stochastická* (*statistická, pravděpodobnostní*) *disambiguace*
2. *pravidly řízená disambiguace*

Stochastická disambiguace

V současné podobě jsou korpusové texty ČNK automaticky disambiguovány programovým vybavením koncipovaným na základě *stochastického modelu*, který je založen především na pravděpodobnostech přechodu mezi jednotlivými značkami v morfologicky analyzovaném (tedy dosud nedisambiguovaném) textu. Princip tohoto typu disambiguace spočívá v tom, že se nejprve ručně (tj. správně) označkuje větší množství textů (o rozsahu řádově několika set tisíc slov, který je pro ruční disambiguaci ještě únosný), a vznikne tak tzv. *trénovací korpus*. Statisticky koncipovaný disambiguační program (tzv. *tagger*) se poté „naučí" toto správné značkování, tj. učiní si představu o pravděpodobnostech přechodu mezi jednotlivými značkami a jejich četnostech, kterou uloží do svých vnitřních tabulek. Program, který se takto naučil správně označkovaný text, poté aplikuje své „znalosti" získané z trénovacího korpusu na dosud nedisambiguovaný korpus a tento korpus s větší či menší úspěšností disambiguuje.

Nejlepší programy pro stochastickou disambiguaci korpusů angličtiny dosahují úspěšnosti zhruba mezi 97 a 98 %, úspěšnost morfologické disambiguace v ČNK stochastickou metodou je zhruba na úrovni 94 % (tzn. že zhruba každé 16. slovo je disambiguováno chybně). Uvedený rozdíl vyplývá zejména z odlišných typologických vlastností češtiny a angličtiny: jazyk s poměrně velmi pevným slovosledem, jakým je angličtina, lze stochastickými metodami založenými na statistické distribuci četnosti posloupností značek zpracovat samozřejmě mnohem úspěšněji. Na rozdíl od angličtiny, kde se typické posloupnosti značek dané pevným slovosledem vyskytují přesvědčivě často, poskytuje čeština jen málo slovosledných záchytných bodů a počet syntaktických konfigurací (tj. slovosledných posloupností o *n* značkách) s přibližně obdobnou četností je v ní podstatně větší; výběr náležitých značek je tu proto nevyhnutelně méně jednoznačný. Korpusové texty se v rámci ČNK dosud disambiguovaly pouze stochasticky, v současné době se však pracuje na vývoji disambiguace řízené pravidly (viz níže).

Hlavní problém, na který narážejí všechny stochasticky koncipované taggery, tkví v nedostatku tzv. trénovacích dat. Syntagmatická a slovosledná variabilita textů je tak velká, že stochastické taggery se prostě nemohou naučit všechny možné posloupnosti značek. Pokud je rozdíl v četnostech různých morfologických interpretací výrazný (např. forma „*se"* se jakožto předložka vyskytuje pouze v 9 % případů a jakožto reflexivní částice nebo reflexivní zájmeno v 91 % případů), je disambiguace ještě poměrně úspěšná; pokud se však četnosti různých interpretací sobě blíží, dochází k výběru nesprávné interpretace mnohem častěji. Zvláštností stochastické disambiguace je také to, že stochastické taggery někdy jasnozřivě „uhodnou" správnou variantu ve složitém kontextu a zároveň se dopustí hrubé chyby v kontextech, kde příslušný výběr správné morfologické interpretace je (relativně) jednoznačný. Příkladem takové hrubé chyby je např. výběr nominativní morfologické interpretace substantiva následujícího v textu korpusu bezprostředně po předložce, což (s výjimkou několika málo přejatých předložek) jazykový systém češtiny zcela vylučuje.

I přes relativně vysoký počet chyb způsobených stochasticky založenými taggery je však přece jen lepší pracovat s částečně chybně označkovaným korpusem než na jakoukoli interpretaci zcela rezignovat. Ukazuje se ovšem, že úspěšnost automatické disambiguace by se dala podstatně zvýšit disambiguací koncipovanou na základě syntaktických pravidel.

Pravidly řízená disambiguace

Vzhledem k tomu, že úspěšnost výše charakterizované stochastické disambiguace českých textů je uspokojivá pouze částečně, byl zahájen vývoj metody disambiguace založené na syntaktických pravidlech. Její podstatou je intuitivní formulace celé řady syntaktických pravidel, která odrážejí syntaktické konfigurace češtiny dané jejím vnitřním systémem. Jakmile je formulováno určité pravidlo, které vyplynulo z analýzy obecné chyby, ihned se počítačově implementuje a ověřuje na datech korpusu. Poněvadž tato metoda modeluje jazykový systém, není – na rozdíl od metody stochastické – závislá na trénovacích datech a vlastně je vůbec nepotřebuje. Pokud je možné formulovat nějaké pravidlo se stoprocentní jistotou, budou i data korpusu značkována správně, pokud ovšem není v textu korpusu chyba. Na chyby v textech (např. chybějící slovo či čárka, nesprávná morfologická analýza aj.) je pravidly řízený tagger velmi citlivý, dokáže však některé takové chyby i odhalit. Jelikož je vývoj této metody dosud na počátku, nelze ještě její úspěšnost exaktně kvantifikovat.

2.4.6 Technické zabezpečení ČNK

Texty přicházejí do ÚČNK v mnoha podobách, které je třeba různými softwarovými nástroji zkonvertovat do jednotného formátu (viz 2.4). Po konverzi jsou data uložena ve velkém počtu různě velkých souborů. Těm je třeba zajistit dostatek prostoru na pevném disku, rozumnou dobu přístupu pro všechny oprávněné uživatele a přijatelnou spolehlivost počítačového systému. Z těchto dat se pro vlastní práci s *korpusovým manažerem* automaticky generují binární soubory, jejichž formát je uzpůsoben rychlému vyhledávání na počítači. Tyto soubory jsou extrémně rozsáhlé a zároveň vyžadují rychlý přístup. Kdykoliv mohou být znovu vytvořeny a proto nemají vysoké nároky na zabezpečení. Pro uložení a zpracování zkonvertovaných dat a jejich binárních verzí je vhodný serverový operační systém. Jak bylo uvedeno výše, data jsou zpřístupněna pro lingvistickou práci pomocí speciální sady programů, nazývané *korpusový manažer*. Jde o program, který musí splňovat dvě základní kritéria, z nichž prvním je dostatečná rychlost při vyhledávání požadovaných lingvistických jevů a druhým uživatelsky příjemné rozhraní. Protože výsledky hledání se obvykle dále zpracovávají, je pro následné zpracování k dispozici další samostatný program. Celý korpusový manažer tak tvoří tři samostatné moduly: nástroje na vyhledávání (implementované v jazycích *C*, *C++* a *Perl*), komunikační program, který zaznamenává výsledky zadaného vyhledání a dále je upravuje, popř. třídí (program je rovněž implementován v jazycích *C*, *C++* a *Perl*), a vlastní uživatelské rozhraní (implementované v jazyku *Tcl/Tk*), které umožňuje zadávat dotazy do korpusu, provádět další operace s daty a zobrazovat výsledky těchto akcí. Všechny tři uvedené programy mohou být spuštěny na jednom počítači, ovšem obvykle jsou první dva spuštěny na serveru a poslední je provozován lokálně na osobním počítači uživatele. Počítače, na kterých se v současné době zpracovává ČNK, se dělí do dvou kategorií: první jsou tzv. *pracovní stanice*, tj. počítače, na kterých pracují jednotliví uživatelé; druhou tvoří *servery* – centrální počítače sloužící všem uživatelům.

Pracovními stanicemi jsou běžná PC pracující pod operačními systémy Windows 2000 nebo Windows 98. Jako servery slouží 4 výkonné počítače založené na platformách Intel a AMD. Jako operační systém byl z důvodu stability a dostupnosti zvolen Unix/Linux. Veřejně přístupným aplikačním serverem je dvouprocesorové PC s procesory Intel Pentium III se 3 GB operační paměti a 3x18GB SCSI disky v rychlém poli RAID 0. Zde jsou uložena

korpusová data, se kterými pracuje korpusový manažer. Hlavním souborovým a výpočetním serverem je AMD AthlonXP s 512MB pamětí a dvěma 160GB SCSI disky uspořádanými v zabezpečeném poli RAID 1. Zálohovací server je vybaven páskovou mechanikou DAT DDS4 a DVD vypalovačkou. Vedlejší výpočetní server slouží primárně pro disambiguaci textu.

2.4.7 Přístup k ČNK

První a nejjednodušší možností, jak si vyzkoušet, co ČNK nabízí, je navštívit internetové stránky ÚČNK na adrese http://ucnk.ff.cuni.cz. Zde je k dispozici veřejný přístup ke korpusu PUBLIC, který má stejné procentuální zastoupení žánrů jako korpus SYN2000, ale je mnohem menší. Oproti velkému korpusu má některá omezení: kromě velikosti – obsahuje „pouze" 30 milionů slovních tvarů – je to omezený kontext, v němž se zobrazuje hledané slovo, a možnost vyhledávat pouze izolované slovní tvary, nikoli skupiny slov.

Užívání tohoto korpusu je velmi jednoduché: slovo (nebo třeba jen příponu nebo předponu) napíšeme do vstupního pole a stiskneme tlačítko *Hledej*. Chceme-li vyhledat různá slova, která mají společnou například příponu, použijeme pro libovolný počet předcházejících písmen řetězce ".*" (tedy: dotaz ".*tel" umožní vyhledání všech slov končících skupinou *-tel*, např.: *nepřítel, pytel, datel, jetel, majitel* atd.). Kontext, ve kterém vyhledané slovo vidíme, je možné rozšířit maximálně na 60 znaků před slovem a 60 znaků za ním. I přes tato svá omezení ukazuje korpus PUBLIC velmi názorně, jaké možnosti nabízí počítačové zpracování jazykového materiálu.

Pro náročnější práci ovšem tento korpus nestačí. Proto nabízíme plný přístup ke stomilionovému reprezentativnímu korpusu SYN2000, s nímž lze dnes pracovat pomocí speciálně vyvinutého sofistikovaného vyhledávacího programu GCQP.

Korpus SYN2000 vznikl z textů, které byly ÚČNK poskytnuty na základě smlouvy o jejich nekomerčním využití (viz 2.4.1). Proto se také každý zájemce o plný přístup ke korpusu zavazuje, že data získaná z korpusu nepoužije ke komerčním účelům. Text prohlášení o nekomerčním používání korpusu, podrobné informace o podmínkách získání přístupu ke korpusu SYN2000 a podrobný návod na instalaci korpusového manažeru GCQP se nachází na adrese http://ucnk.ff.cuni.cz/manual.

3. Využití korpusů

Pro svou obsažnost a univerzální povahu jsou velké korpusy (zvláště korpusy obecně reprezentativní, k nimž patří i SYN2000) neocenitelnými zdroji informací, a to nejen pro lingvisty, ale také pro odborníky z řady dalších oborů, zejména literární vědce, informatiky, sociology, psychology a pedagogy. Obecně vzato existuje jen málo oborů, které lze studovat primárně jinak než skrze jazyk, a možnosti, jak využít rozsáhlý soubor textů, reprezentující nejen tento jazyk, ale i hodnoty, problémy a zájmy jeho uživatelů, jsou proto velmi široké. Korpus však není určen pouze odborníkům: první zkušenosti naznačují potěšitelnou skutečnost, že k jeho častým uživatelům budou patřit studenti a že živý zájem o něj jeví i zainteresovaní laikové.

Zaměřme se však na skupinu nejčastějších uživatelů, na lingvisty a studenty jazyka. Ti dnes využívají korpusových dat především pro výstavbu popisů jazyka a pro tvorbu a ověřování teorií, stále častěji však i k tvorbě velkých aplikací založených na korpusových datech (k nim patří především už zmínění lexikografové). Co vlastně nabízí korpus těmto uživatelům navíc ve srovnání s tím, na co byli zvyklí z tradiční excerpce? Sama konkordance (tj. výpis všech řádek s výskytem hledaného jevu v kontextu), která je základním výsledkem hledání v korpusu, jim především předkládá data podstatně lépe podložená a zasazená do dostatečného, libovolně rozšiřitelného kontextu; vedle toho jim korpusový program poskytuje i nejdůležitější frekvenční a statistické charakteristiky příslušných dat a další techniky průzkumu souvislostí mezi slovy. Na základě analýzy všech těchto údajů mohou lingvisté velmi snadno získat představu o tom, co a jak se používá typicky a co okrajově (zaměření na typičnost užití je samozřejmě mimořádně důležité pro lexikografii, ale zdaleka nejen pro ni). Máme-li navíc možnost porovnat získané výsledky se staršími daty např. v diachronním korpusu nebo ve starších gramatikách či slovnících, můžeme také jednoduchou extrapolací činit závěry o pravděpodobných směrech vývoje, vývojových tendencích apod.

Mimo typičnost jednotlivých jevů lze analýzou výsledků hledání v korpusu dobře ukázat i **škálovitou povahu** většiny jevů v jazyce, tedy jejich přechodnost a neostrost hranic mezi nimi. Proto je korpus tak důležitý pro studium přirozené a všudypřítomné jazykové **variability**, a to i v oblastech, kde se jí společnost umělými příkazy (např. v podobě nadiktovaných pravidel pravopisu, neopírajících se o znalosti dat) brání. Korpus nepochybně nabízí mnohem více druhů informací, než jsme naznačili a než si v současné době jsme schopni vůbec uvědomit; některá využití korpusu na své uživatele jistě teprve čekají.

4. Závěr – budoucnost ČNK

Na základě ČNK vzniknou nové popisy češtiny

– nová mluvnice, výkladový slovník, další typy speciálních slovníků, jazykovědné studie, různé příručky a učebnice. Konkrétním výstupem korpusu SYN2000 je mj. také *Frekvenční slovník psané češtiny*, který byl již předán Nakladatelství Lidové noviny k publikaci. Před dokončením je frekvenční slovník pražské mluvené češtiny, který vznikl na základě PMK. Chystá se příručka cvičení a úkolů pro využití ČNK studenty. Před dokončením je nový manažer Bonito.

ČNK je chápán jako projekt kontinuální, a proto v dalších letech bude ÚČNK zpřístupňovat další korpusy a informace v nich tak, aby v roce 2011 bylo dosaženo jedné miliardy slovních tvarů. Dále bude probíhat sběr mluvených textů tak, aby byly v nejbližší možné době pokryty všechny oblasti České republiky.

Souběžně také probíhá studium některých jazykových jevů, pro které je ideální právě využití korpusu

– např. jazyková variabilita, spojitelnost slov. Rádi bychom se věnovali slovní zásobě období totality a lákají nás autorské slovníky předních českých autorů.

V ČNK je možné najít i terminologii. ČNK je ovšem sestaven tak, aby uživatel získal obecné informace o jazyce. Dozvíme se z něj tedy především to, který termín proniká do obecného

užití. Pro terminologické slovníky bude ovšem nutné sestavovat speciální korpusy, které zahrnou pouze texty daného oboru.

Chystáme se dále ve spolupráci s kolegy ostatních lingvistických pracovišť FF UK a se zahraničními spolupracovníky vytvořit více než dvacet paralelních korpusů různých jazyků spolu s češtinou, které poskytnou ideální studijní materiál pro studenty i vyučující, ale i materiál pro vznik nových překladových slovníků.

ÚČNK měl to štěstí, že začal svou práci včas (záhy po roce 1989), a tím se mu podařilo vytvořit korpus (slovanského jazyka) srovnatelný ve velikosti a zpracováním s korpusy anglickými. Čeština, jakožto flektivní jazyk, přináší zpracovatelům při lingvistickém značkování ovšem nesnadno řešitelné problémy. I s nimi je nutné se postupně vyrovnat.

ČNK je určen především lingvistům. To ovšem neznamená, že ho neocení např. i psychologové, sociologové, informatici. V korpusu je možno hledat, kromě lingvistických dotazů, i informace encyklopedické. Lze si na něm ověřit své pochybnosti o konkrétních jazykových jevech.

Korpus se ovšem ne vždy shoduje s pravidly pravopisu, protože odráží skutečné jazykové užití. Lze se v něm informovat i o významech nových slov, který vyplývá z větného kontextu.

V České republice vznikají jazykové korpusy i se speciálním technickým zaměřením. Reprezentativní korpus, který je určený k obecnému lingvistickému využití a je kompatibilní s korpusy ostatních jazyků, však pro češtinu a české prostředí existuje jediný. Je jím právě Český národní korpus, který češtině otevírá dveře nejen do Evropy, ale do celého světa.

Příloha: Ukázky využití ČNK

I.

Otázka: Kdo řekl citát? „Veškeré kvaltování toliko pro hovada dobré jest.“

Odpověď najdete v ukázkách.

-Ještě k něčemu nás Jan Amos Komenský - učitel národů - inspiroval nejen proto, že prý : " všeliké <kvaltování> toliko pro hovada dobré jest ".

-Jsme kulturní národ Jana Ámose Komenského a dobře víme, že veškeré <kvaltování> toliko pro hovado dobré jest.

-Veškeré <kvaltování> toliko pro hovada vhodné jest, řekl J. A. Komenský, a to byla nějaká osobnost!

II.

Jak mám psát následující slovo: ***jazz*** nebo ***džez***?

Odpověď: podle počtů výskytů v ČNK většina uživatelů dává přednost původnímu anglickému psaní.

džez – celkem 59krát

ukázka deseti výskytů

ještě dozpíval : <Džez> je dnes život můj

toho měla pravdu , <džez> skutečně miloval

zajedu poslechnout dobrý <džez>

a pak chtěli hrát <džez> na varhany

přesedlala na <džez>

na plné pecky pustila si <džez>

Mě baví <džez> džez džez . . .

Chvalte Hospodina <džezem> , blues i symfonií

po konzervatoři se věnoval <džezu>

Zemánkovou , protagonistkou <džezu> čtyřicátých let

jazz – celkem 1833krát

ukázka deseti výskytů

kvůli té holce vytrpěl , a na <jazz> úplně kašlala

zaměřená na vážnou hudbu , uvítá <jazz> koncertní sérií Spirituals

stěžejním osobnostem newyorského <jazzu>

Kolem pojmu <jazz> je hodně nedorozumění

protože ve Varech zájem o <jazz> velký nebyl

Pro poslech klasiky a <jazzu> se tyto přístroje sotva hodí

Brom , který často vychází z <jazzu> a jazzových improvizací .

spojuje prvky funky , acid <jazzu> , rocku a blues

do světa latinského soul <jazzu> , v němž posluchač pozná

mezi volně improvizovaným <jazzem> a groovem acid jazzu

III.

Otázka: Co znamená slovo ***píár***?

Odpověď najdete v ukázkách.

*-Třetí důležité poradě tu říkají " **<píár>** ", což má skutečně původ v anglickém " public relations " neboli styk s veřejností.*

*-Na " **<píár>** " poradě se tým ODA například rozhodl, že bude potřeba říkat více věcí voličům přímo na mítincích, a naopak trochu rezignovat na tiskové konference, o jejichž obsahu nejsou novináři povinni zveřejnit ani řádku.-Píár se tomu teď moderně říká a firmy na tohle **<píár>** teď dávají stohy bankovek.*

*-Když krachuje banka, která stojí a padá s jménem svých akcionářů, **<píár>** poradci, pud sebezáchovy, ale hlavně zdravý rozum a etika chybějí.*

Literatura

1. BURNARD, L. *Users' Reference Guide for the British National Corpus*. Oxford : Oxford University Press, 1995.
2. ČERMÁK, F. Jazykový korpus : prostředek a zdroj poznání. *Slovo a slovesnost, 1995, roč.* 56, č. 2, s. 119-140.
3. ČERMÁK, F. Czech National Corpus : a case in many contexts. *International Journal of Corpus Linguistics*, 1997, č. 2, s. 181-197.
4. ČERMÁK, F. Czech National Corpus : its character, goal and background. In *Text, speech, dialogue : proceedings of the first workshop on text, speech, sialogue – TSD '98 : Brno, Czech Republic, September 23-26, 1998*. Brno : Masarykova Universita, 1998, s. 9-14.
5. ČERMÁK, F. Language corpora : the Czech case. In *Text, speech and dialogue : proceedings of the fourth international conference TSD 2001 : Železná Ruda, Czech Republic, September 11-13, 2001. Berlin : Springer, 2001.*
6. ČERMÁK, F., KRÁLÍK, J., KUČERA, K. Recepce současné češtiny a reprezentativnost korpusu. *Slovo a slovesnost,* 1997, roč. 58, č. 2, s. 118-124.
7. KOCEK J., KOPŘIVOVÁ, M., KUČERA, K. (eds.). *Český národní korpus : úvod a příručka uživatele* Praha : Filozofická fakulta UK, 2000.
8. HALLIDAY, M. A. K. *Spoken and written language*. 2nd ed. Oxford : Oxford University Press, 1989. 109 s.
9. HLAVÁČOVÁ, J., RYCHLÝ, P. Dispersion of words in a language corpus. In *Text, speech and dialogue : second international workshop, TSD '99, Plzen, Czech Republic, September 13-17, 1999 : proceedings*. Berlin : Springer, 1999.
10. KRUYT, J. G. *Design criteria for corpora construction in the framework of a European corpora network. Final report.* Leiden : Institute for Dutch Lexicology, 1993.

11. KUČERA, K. Diachronní složka Českého národního korpusu : obecné zásady, kontext a současný stav. *Listy filologické*, 1998, roč. 121, s. 303-313.
12. NORLING-CHRISTENSEN, O. Preparing a text corpus : computational tools and methods for standardizing, tagging and structuring text data. In KIEFER, R. et al. (eds.). *Papers in Computational Lexicography, COMPLEX '92.* Budapest : Research Institute for Linguistics, Hungarian Academy of Science, 1992, s. 251-259.
13. NUSBAUM, H. C. A stochastic account of the relationship between lexical density and word frequency. In *Research on speech perception*, Indiana University, 1985.
14. PETKEVIČ, V. Neprojektivní konstrukce v češtině z hlediska automatické morfologické disambiguace. In Hladká, Z., Karlík, P. (eds.). *Čeština - univerzália a specifika* 3. Brno : Masarykova univerzita, 2001, s. 197-206.
15. ŠULC, M. *Korpusová lingvistika : první vstup.* Praha : Karolinum, 1999. 94 s.