Jarmila Panevová (Univerzita Karlova, Praha)

Čím může bohemistice přispět současná počítačová lingvistika?

I. Po několik destiletí (od 2. poloviny 20. století) se mohlo zdát, že počítačová a formální lingvistika (třeba i pod názvem matematická lingvistika) přispívá v rámci lingvistiky jen k jejím teoretickým aspektům, a to přesnějšími, exaktně formulovanými tvrzeními opřenými o jasná kritéria. Ve svých aplikacích směřovala matematická (počítačová) lingvistika k vytváření systémů strojového překladu, automatického ukládání a vyhledávání textových informací nebo k tvorbě dotazovacích systémů k různě strukturovaným databázím na základě použití přirozeného jazyka. Poznání vnitřní struktury těchto systémů je nezávislé na jejích uživatelích, a pokud systém poskytuje žádoucí výsledky, nemusí vnitřní struktura uživatele vůbec zajímat. Lingvistické a bohemistické výsledky dosažené teoretickými a komputačními lingvisty zůstávaly mnohdy nepovšimnuty a v lingvistice nevyužity se zjednodušeným odůvodněním, že přece směřují k počítačovým aplikacím, tj. že jsou formulovány pro počítače.

Od 90. let 20. století vznikem korpusové lingvistiky jako odvětví lingvistiky počítačové se situace prudce změnila. Vytvářejí se obrovské elektronické textové korpusy (srov. např. Čermák, 1995, Čermák ad. (red.), 2000) a vznikají i tzv. označkované (anotované) korpusy, které obsahují morfologické, popř. i syntaktické (nebo i další) informace, ať už jsou do anotací vloženy manuálně nebo automaticky. Můžeme na tomto místě zcela po právu konstatovat, že čeština je z tohoto hlediska zpracována na takové úrovni, že velmi dobře obstojí i v porovnání s korpusy jazyků rozšířenějších a důkladněji popsaných (jako je např. francouzština nebo němčina). Korpusová lingvistika dnes nabízí možnost širokého využití počítačových korpusů jako materiálové základny pro všechny bohemisty, a tedy i pro učitele a studenty Slezské univerzity, kteří se, jak je autorce dobře známo, o těchto možnostech v rámci studia bohemistiky dozvídají. Bohemisté čím dále tím více oceňují nesporné výhody, které pro studium, popis i výuku současné češtiny má možnost rychlého počítačového vyhledávání v obrovských korpusech jazykových dat. Aby tato data byla opravdu lingvistické veřejnosti dostupná, je třeba do jejich sbírání, ukládání a formátování uložit velké množství lidské práce. Informatici a programátoři vytvářejí programové nástroje, umožňující s těmito daty co nejefektivněji a s minimálním zaškolením pracovat. Uživatelé Českého národního korpusu znají korpusový manažer (Kocek ad. (red.),

2000) nástroj, v němž uživatel formuluje své dotazy. Lingvisté ve spolupráci s informatiky vytvářejí systémy značkování, přidělování gramatických informací slovním tvarům v jazykových korpusech. Pro stručnost zde necháme stranou otázky, zda značkovací procedury jsou automatické, poloautomatické nebo manuální, a odkážeme zde čtenáře na jiné zdroje (např. Hajič, 1998, Hajič ad., 1998). Zde se pokusíme ilustrovat, jak lze základních korpusových zdrojů, které dnes pro češtinu existují, využívat a jaké možnosti pro výzkum současného jazyka poskytují, a to na základě několika příkladů.

V ČNK (vyvíjeném v Ústavu Českého národního korpusu na Filozofické fakultě UK v Praze) může zájemce vyhledávat v lineárním českém textu nejen výskyty jistého slovního tvaru, ale díky programovému prostředku, který automaticky provádí lemmatizaci, výskyt tohoto slova ve všech jeho tvarech, nebo jen v těch tvarech, které uživatele zajímají, v kontextech tak širokých, jak si uživatel v dotazu zadá. Žádným problémem pak není ani vyhledání jistých kolokací (jednotlivé složky kolokace mohou/musí být bezprostředně vedle sebe v textu, nebo mohou být ve větě odděleny, tento požadavek si uživatel rovněž ve svém dotazu nastaví), např. *hrubý domácí produkt, uzavřít… smlouvu… o spolupráci* atd.

Na Matematicko-fyzikální fakultě UK v Praze (v jejím Ústavu formální a aplikované lingvistiky a Centru komputační lingvistiky) se část ČNK podrobuje značkování na třech rovinách (morfologické, na rovině zvané analytická, která je blízká rovině povrchové syntaxe, a na rovině tektogramatické, tj. na rovině tzv. „hloubkové“ syntaxe). Tím se vytváří **označkovaný korpus** zvaný **Prague Dependency Treebank** (PDT, Pražský závislostní korpus). Protože morfologická úroveň je již nepřímo zahrnuta v ČNK prostřednictvím modulu lemmatizace, nebudeme se o ní zde dále šířit. Značkování na tektogramatické rovině je značně náročné lingvisticky i metodologicky (jak programovými nástroji manuální značkování anotátorům co nejvíce usnadnit a jak zajistit konsistenci anotování); je proto zatím v počátcích a jeho výsledky (cca 4 000 vět) nejsou dosud běžnému uživateli dostupné; jde ostatně o vzorek textu ne dost reprezentativní pro vyhledávání jevů zachycených v tektogramatickém stromě . Proto ani o něm zde blíže mluvit nebudeme.

Na **analytické rovině** je zpracováno cca **97 000 vět,** které byly vybrány z ČNK tak, aby jistým způsobem reprezentovaly žánrové složení ČNK. Ty jsou zachyceny v PDT v podobě závislostního stromu s ohodnocenými uzly. Ohodnocení, které uzlům přidělují anotátoři, odpovídá v podstatě tradičním větným členům (dílčí odchylky jsou dány konvencemi pro zachycení koordinovaných a aponovaných struktur, vsuvek a dalšími dílčími konvencemi vyvolanými technickou stránkou anotování, popis těchto konvencí je obsažen např. u Hajiče, 1998; jde zejména o fakt, že každý tvar nebo symbol mezi dvěma mezerami je

třeba zachytit jako samostatný uzel v analytickém stromě, tzn. že nejen autosémantická, ale i synsémantická slova a interpunkční a další znaky tam mají svůj samostatný uzel). Anotátor kontroluje strukturu, která byla automaticky předzpracována a je mu na obrazovce nabídnuta, opraví a doplní údaje, v nichž se automatický modul (tzv. pre-processing) dopustil chyby. Tak jako pro ČNK je běžnému uživateli k dispozici korpusový manažer, jsou pro prohlížení stromů a vyhledávání podstromů podle uživatelova dotazu určeny programové nástroje TRED a NETGRAPH 2.0, které jsou zájemcům k dispozici na uvedených pracovištích MFF UK.

II. Podívejme se teď na některé konkrétní gramatické problémy, jak se nám budou jevit při vyhledávání v ČNK a v PDT.

(1) Při studiu funkcí české předložkové vazby *za + Akuzativ* jsme z ČNK (v jeho verzi SYN2000) při formulaci dotazu na výskyt tohoto předložkového pádu s dodatečnou podmínkou, že mezi předložkou a jménem v příslušném pádě mohou stát nanejvýš 3 slovní tvary, získali 213 779 dokladů. Při letmém náhledu zjišťujeme, že ve vybraném materiálu značně převažuje význam temporální (*Za chvíli určitě narazí na střelce…, Nedá se zvládnout za jeden rok…)*, dost zastoupen je význam směrový (*Skočil za roh).*Okrajová tu nejsou ani frazeologizovaná užití (*mít/vzít si za manželku, popadat se za břicho).* V příkladech jako *vzít za ruku, zatahat za košili, přivést někoho za ucho* se užití tohoto předložkového pádu pokládá za určení prostředku (Šmilauer, 1966, s. 302), což nepokládáme za zcela nesporné (srov. *zatahal ho levou rukou za košili,* kde jde o dva navzájem nezaměnitelné „prostředky“, vyžadující zřejmě jemnější významovou diferenciaci; tu zde necháváme otevřenu další diskusi). Z ČNK jsme samozřejmé získali i materiál, v němž je příslušný předložkový výraz v adnominální platnosti. Protože nás ale zajímala především funkce *za + Akuzativ* v platnosti objektu (patientu nebo fektu v termínech hloubkové syntaxe), tedy případy, kde jde o pád vazebný (rekční), je nutno získaný rozsáhlý materiál nějak omezit. Můžeme tady postupovat dvěma způsoby: zaměřit svůj dotaz pouze na ta slovesa v ČNK, u nichž přítomnost objektu v této formě předpokládáme, jako např. *(vy)měnit, (za)platit, dát, pokládat, považovat, (z)volit,(po)děkovat* apod., čímž ovšem nezaručíme úplnost excerpovaných dokladů, ale jenom potvrdíme hypotézu o valenci u sloves, která jsme předem vytipovali. Druhý způsob je použití PDT, tj. korpusu sice objemově menšího, ale strukturovaného a opatřeného analytickými funkcemi (větnými členy), z nichž nás právě jedna (Obj) zajímá. Po prohledání PDT v současné verzi (97 000 vět) jsme získali 883 výskytů objektového *za + Akuzativ* tak, že jsme vyhledávali podstrom, v němž na předložce (s analytickou funkcí *AuxP*) závisí

substantivum, které má ve své morfologické značce *Akuzativ* a zároveň má analytickou funkci *Obj* (výskyt v koordinaci byl počítán za jediný výskyt). Tento vzorek obsahuje přesně ty výskyty, o které se zajímáme, a počet výskytů je manuálními silami dobře zvládnutelný (např. *…kterého považoval za svého přítele, … přednosta stanice odpovídal za pohyby desítek vláčků).*

(2) Při studiu sémantiky předložkového pádu *o + Akuzativ* jsme na základě konstrukce *píchla se o trn* zjišťovali sémantický rozdíl mezí (i*) píchl se o nůžky* a (ii) *píchl se nůžkami.* Jejich zjevný sémantický rozdíl ukazuje (stejně jako rozdíl *zatahal ho za levou ruku/ za košili* a *zatahal ho levou rukou* zmíněný v bodě 1 výše), že nejde o sémanticky stejný typ prostředku, ale že je potřebná jemnější klasifikace (srov. k tomu i Hajičová ad., 2002, s. 168). O určení prostředku/nástroje lze jednoznačně mluvit u (ii), v konstrukci (i) jde o nějaký podtyp prostředku, ale „nástrojový" předmět (v tomto případě *nůžky*) se nepohybuje, je statický. Rozdíl mezi (i) a (ii) spočívá i v tom, že v (ii) může jít o činnost úmyslnou, v (i) je v každém případě neúmyslná. Zatím pouze u slovesa *zavadit* jsme shledali, že tento „nepohyblivý nástroj" (který jsme pracovně nazvali **překážka**) se jeví jako obligatorní (základový člen větné struktury). Při vyhledání této konstrukce v ČNK (vyhledání provedla M. Lopatková pro účely tvorby valenčního slovníku jako plánované součásti PDT) jsme získali další případy souvýskytu prostého instrumentálu a předložkové vazby *o + Akuzativ* jako např. *bouchla se hlavou o pelest, vůz se otřel nárazníkem o skálu, praštil se hlavou o okno, zarazil se rukama o zeď, opřel se pohodlně zády o opěradlo*. Tyto příklady rovněž dokládají, že každé z obou „nástrojových" substantiv vyjadřuje jiný sémantický vztah ke slovesu a že forma *o + Akuzativ* má opravdu sémantickou povahu překážky, která subjektu nebo objektu buď slouží jako opora, nebo mu způsobuje újmu.

(3) Při studiu jistého typu rozvití přívlastkového přídavného jména, v němž často vznikají tzv. „falešné větné dvojice" (jako v příkladu *Dívka rovná ve výloze vystavený kabát*), jsme vyslovili hypotézu, že tato „citlivá" místa ve větě vznikají zejména, je-li tento přívlastek preponován svému řídícímu substantivu. V některých takových konstrukcích sice syntaktická homonymie nevzniká, ale přesun preponovaného rozvitého přívlastku do postpozice by strukturu věty zprůhlednil a snad i stylisticky zlepšil (podrobněji k této hypotéze viz Panevová – Ribarov, v tisku). Když jsme se snažili vybrat takové konstrukce v ČNK, kde jsme museli svůj požadavek formulovat v podobě dotazu skládajícího se z lineárního pořadí slovních druhů a jejich tvarů (bylo zadáno pořadí „sloveso – předložka – (až o 2 pozice vpravo vzdálené) substantivum – adjektivum"), získali jsme **212 364 dokladů**. Abychom toto množství zredukovali, dodali jsme do dotazu omezení, že adjektivum má být

deverbativní. Tím jsme získali podmnožinu minulého souboru dokladů čítající po této redukci **21 316 dokladů**, což je asi 10% původního vzorku. Dotazem s omezovací podmínkou na **deverbativní** adjektivum a **jeho předložkové rozvití** jsme ovšem vědomě rezignovali na případy jako (1) a (2), kterých jsme posléze našli v PDT pouze několik:

(1) Většina lázeňských společností nabízí **pro obchodování nezajímavá** množství kapitálu *(#40 pdt/public/analytic/ca11am.fs)*

(2) Výrazně brzdící roli zde hrají **zahraniční konkurencí ovlivňované** ceny vývozu mimo SR *(#38 v208_2vzberb.fs)*

I tak přirozeně mnoho dokladů z ČNK nevyhovovalo hledané syntaktické struktuře a musely být ručně vyloučeny, jako např. (3) a (4):

(3) Barney < ležel na podlaze obývacího pokoje >

(4) < vytáhl z kapsy pomačkanou krabičku > anglických cigaret

Detailní analýza relevantních dokladů, kterých bylo ve vzorku 430 náhodně vybraných příkladů (z 21 316 celkových, viz výše) pouze 50, je obsažena ve stati Panevová – Ribarov (v tisku). V ní je zhodnoceno, do jaké míry odpovídá materiál vyslovené hypotéze (o výskytu „falešných větných dvojic“ a možnosti přesunout celé adjektivní rozvití do postpozice).

Protože nás zajímala určitá syntaktická konstrukce, sáhli jsme pochopitelně i po materiálu z PDT, kde jsme pomocí nástroje NETGRAPH 2.0 vyhledávali podstromy, v nichž bylo substantivum rozvito adjektivem, které je samo rozvito (A) předložkovým pádem substantiva, (B) prostým pádem substantiva. Na základě vyhledávání (A) jsme získali **2 701 podstromů,** na základě (B) pak **2 077 podstromů**. Na adjektivum tu nebyla kladena žádná další podmínka, což znamenalo, že byl zahrnut i typ příkladů, jako je v (1), kde nejde o adjektivum deverbativní. Ani rozvití adjektiva nebylo omezováno další podmínkou, vyhledala se tedy předložková i bezpředložková rozvití objektová i příslovečná. Současná verze NETGRAPHU neumožňuje bohužel zatím rozlišit rozvití zleva a zprava, ačkoli analytické stromy (sloužící jako data pro vyhledávání) tuto informaci obsahují (v podobě uspořádání uzlů ve stromě zleva doprava). I tady bylo tedy třeba získaný materiál ručně přetřídit a eliminovat rozvití zprava. Při detailní analýze 10% příkladů ze vzorku (A) a (B) bylo z 270 výskytů v souboru (A) 213 dokladů pro náš výzkum nerelevantních, v souboru (B) jich bylo 153 z 208 nerelevantních. Tím se potvrdil i předpoklad, že postponovaný rozvitý adjektivní

přívlastek je častější a že tak autoři textů dávají přednost přehlednější syntaktické struktuře. Z relevantních dokladů na dotaz (A) s předložkovým rozvitím a na dotaz (B) s nepředložkovým rozvitím uvedeme alespoň několik příkladů: (5), (6) ze souboru (A) a (7), (8) ze souboru (B):

(5) Kdo by si třeba v roce 1898 pomyslel, že **za pouhých 600 franků neprodané Cézannovo zátiší**…, bude za dvě desetiletí ceněno na 300 000 franků *(#38 v270_1vyberc.fs)*

(6) … které **za vyhláškou stanovených okolností** mohou povolat tlumočníka nebo překladatele *(#7 v270_5vyberb.fs)*

(7) Růst počtu **teplem zásobovaných bytů** přitom činil jen několik procent *(#52 v208_2.fs)*

(8) Již dříve **vládou vyhlášený Program** rozvoje cestovního ruchu uvedené problémy zcela neřeší *(#17 v208_2vyberb.fs)*

III. Chtěli jsme v tomto příspěvku poukázat na několik aktuálních bodů spojených s tvorbou jazykových korpusů: Korpusová lingvistika nepřináší nové možnosti jen pro výzkum slovní zásoby a pro lexikografii, ale i pro výzkum gramatických jevů. Čím více informací o jazykových jevech do korpusů uložíme, tím více z nich uživatel může těžit. Nechtěli jsme ovšem zakrývat ani fakt, že i tak zbývá pro lingvistu velké množství ruční a intelektuální práce s materiálem, který mu korpus poskytne, ať je jakkoli sofistikovaný. Popis a výklad jazykových jevů zůstane i nadále doménou člověka.

Literatura

Čermák, Fr. (1995): Jazykový korpus: Prostředek a zdroj poznání. *Slovo a slovesnost 56, 119-140.*

Čermák, Fr. – Klímová, J. – Petkevič, Vl. (red.) (2000): Studie z korpusové lingvistiky. *AUC – Philologica 3-4, Karolinum, Univerzita Karlova, Praha.*

Hajič, J. (1998): Building a syntactically annotated corpus: The Prague Dependency Treebank. *In*: Valency and Meaning. Karolinum, Praha, 106-133.

Hajič, J. – Hajičová, E. – Panevová, J. – Sgall, P. (1998): Syntax v Českém národním korpusu. *Slovo a slovesnost 59, 168-177.*

Hajičová, E. – Panevová, J. – Sgall, P. (2002): K nové úrovni bohemistické práce: Využití anotovaného korpusu. Část I. *Slovo a slovesnost 63, 161-177.*

Kocek, J. – Kopřivová, M. – Kučera, K. (red.) (2000): Český národní korpus. Úvod a příručka uživatele. *FF UK, Praha*.

Panevová, J. – Ribarov, K. (v tisku): Za poleznostta na elektronskite jazični korpusi (vrz primerot na eden tip na imenskata fraza vo češkiot jazik. *In:* Festchrift za prof. Zuzanna Topolińska. Institut za makedonski jazik. Skopje.

Šmilauer, Vl. (1966): Novočeská skladba. 2. vydání, SPN Praha.

Jarmila Panevová
Ústav formální a teoretické lingvistiky
Univerzita Karlova
Matematicko-fyzikální fakulta
Malostranské n. 25
118 00 Praha 1
E-mail: panevova@ufal.mff.cuni.cz