Hladká, Zdeňka

Zkušenost s tvorbou korpusů češtiny v ÚČJ FF MU v Brně

Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná. 2005, vol. 54, iss. A53, pp. [115]-124

ISBN 80-210-3705-9
ISSN 0231-7567

Stable URL (handle): https://hdl.handle.net/11222.digilib/101736
Access Date: 30. 11. 2024
Version: 20220831

SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 53, 2005 — LINGUISTICA BRUNENSIA

ZDEŇKA HLADKÁ

ZKUŠENOSTI S TVORBOU KORPUSŮ ČEŠTINY V ÚČJ FF MU V BRNĚ

Podkladem pro následující text byla přednáška v pobočce Jazykovědného sdružení ČR na Filozofické fakultě UK v Praze (9. 12. 2004).

Úvod

Práce s elektronickými korpusy umožňujícími rychlé vyhledávání a statistické hodnocení velkého množství autentických jazykových dat se v posledních letech stává samozřejmostí i v oblasti zkoumání slovanských jazyků. V českém prostředí je institucionálním centrem korpusové lingvistiky Ústav Českého národního korpusu založený v r. 1994 a fakticky fungující na Filozofické fakultě Karlovy univerzity od října 1996. Na tomto pracovišti vznikl *Český národní korpus* (ČNK), jehož zatím nejdůležitější částí je reprezentativní synchronní korpus SYN2000 obsahující cca 100 milionů slovních výskytů. Shromažďuje psané, tj. v zásadě spisovné texty ve stylových proporcích 60 % publicistiky, 25 % odborné literatury a 15 % beletrie.

Snahou pražského korpusového centra je posílit reprezentativnost ČNK rozšířením pestrosti jeho skladby, a to zachycením mluveného jazyka a zařazením perifernějiších typů textů do korpusu jazyka psaného. Do tohoto úsilí se zapojil i Ústav českého jazyka Filozofické fakulty Masarykovy univerzity: v počáteční fázi vytvořením **korpusu běžné mluvy města Brna**, v současné době přípravou **korpusu soukromé korespondence**. První z uvedených korpusů v následujícím příspěvku zmíníme pouze stručně, druhému, který je ve stadiu přípravy a testování, se budeme věnovat detailněji.

1. Korpusové zpracování mluveného jazyka

Význam studia mluveného jazyka, který je z hlediska fylogenetického i ontogenetického primární formou jazykové komunikace a v reálu se na ní podílí zhruba devadesáti procenty, není třeba připomínat. Vytváření přístupných refe-

renčních zdrojů pro toto studium je však obtížné a časově velmi náročné. S problémem zachycení mluveného jazyka se potýká i většina elektronických korpusů (např. *British National Corpus*, který je jakýmsi korpusovým vzorem, obsahuje pouze 10 % mluvených textů).

1.1 Sběr materiálu

Pro korpusové účely je nejprve nutno mluvený jazyk zaznamenat v co nejpestřejší škále situací, a to tak, aby nahrávky byly technicky kvalitní a zároveň přinášely dostatečně spontánní, nestylizovaný projev. Optimální by bylo pořizovat nahrávky tajné, to však naráží na překážky legislativní a morální. Je tedy třeba volit různé kompromisní postupy, např. tajně nahrávat mluvčí z okruhu příbuzných a známých a následně od nich získat svolení ke korpusovému zveřejnění promluv. Nebo alespoň pořizovat nahrávky s fingovaným účelem (např. jako sociologický průzkum), aby mluvčí nevěnovali příliš pozornosti jazykové stránce svého projevu. Při nahrávání je třeba brát v úvahu i to, že záznam rozhovoru více než dvou mluvčích přináší komplikace při identifikaci a přepisu jednotlivých replik.

1.2 Přepis

Náročnou prací je také přepis nahrávek. Elektronické nástroje, které by byly schopny přesně analyzovat zvukový záznam promluv od mnoha různých mluvčích, zatím nejsou k dispozici. Přepisy je tedy nutno pořizovat ručně. Důležitá je přitom volba transkripčních pravidel. V podstatě jde o rozhodování, zda korpusovou podobu co nejvíce přiblížit přesnému fonetickému znění, nebo naopak pravopisným konvencím užívaným v psaných spisovných textech. První způsob je přesnější a bližší realitě, druhý je čitelnější pro běžného uživatele korpusů a zároveň přístupnější pro korpusové nástroje, které jsou zatím primárně vytvářeny pro analýzu jazyka psaného (převážně spisovného), a jsou tedy „přivyklé" jeho pravopisné i grafické podobě. Dosavadní pokusy přepisu mluveného jazyka v elektronických korpusech češtiny (např. v PMK a BMK, viz dále) kompromisně spojují obě uvedené možnosti. Zčásti se řídí pravopisnými konvencemi, např. v reflexi znělostní asimilace, v psaní *i/y*, *ě* apod., zčásti uchovávají realitu projevu, např. při záznamu zjednodušené výslovnosti (*du, sem, prože, šesnác, srce*); k problematickým otázkám patří mj. reflexe splývavé výslovnosti na hranici slov, na jejíž zachycení se většinou rezignuje, protože by komplikovalo delimitaci lexikálních jednotek.

1.3 Morfologické značkování

Dalším nelehkým úkolem při tvorbě korpusů, jejichž účelem je přinášet co nejvíce informací o jazyce, je lingvistické značkování shromážděného materiálu. Zatím nejčastější je značkování morfologické. Ruční značkování by bylo časově neúnosné, proto snahy korpusové lingvistiky směřují k vytváření nástrojů umožňujících alespoň částečnou automatickou morfologickou analýzu. Čeština je sice

v tomto směru vzhledem k velké míře homonymie jazykem značně problematickým, v oblasti automatické morfologické analýzy a následné disambiguace psaných spisovných textů se však už podařilo dosáhnout značných úspěchů. Značkování korpusů mluveného jazyka je ale mnohem obtížnější, a to i v případě, že je přepis nahrávek v co největší míře přizpůsoben pravopisné konvenci, na niž jsou analyzátory „zvyklé". Běžně mluvený jazyk je totiž složitým konglomerátem různých jazykových útvarů a vrstev, má uvolněnou stavbu (takže např. při desambiguaci, tj. zjednoznačňování homonymních forem, v podstatě nelze užít pomocných syntaktických pravidel), objevují se v něm různá komolení, nedořečení, opakování apod. Tyto rysy komplikují automatické rozpoznávání slovních forem i určování kritérií pro jejich přiřazování k lexikálním lemmatům (podrobněji o problematice lemmatizace – i když zejména ve vztahu ke korpusům psaného jazyka – např. Petkevič, 2004). Morfologické značkování korpusů mluveného jazyka vyžaduje speciální úpravu nástrojů vytvořených pro analýzu psaných textů a zároveň mnohem větší podíl ruční práce.

Z výše naznačených a z řady dalších důvodů je vytváření korpusů mluveného jazyka v českém prostředí (ale i v zahraničí) zatím v úplných začátcích. Také korpus brněnské mluvy, který následně stručně představíme, je nutno vzhledem k jeho velikosti a řadě metodologických nedokonalostí považovat pouze za cvičný pokus v dané oblasti.

1.4 Brněnský mluvený korpus

Tzv. *Brněnský mluvený korpus* (BMK, resp. ORAL-BMK) je od r. 2002 veřejně přístupný jako součást ČNK (bližší informace http://ucnk.ff.cuni.cz). Vznikl výběrem materiálu z rozsáhlejšího pracovního souboru nahrávek a přepisů mluvené češtiny města Brna vytvářeného v 90. letech 20. století v ÚČJ FF MU. BMK obsahuje elektronický přepis 250 magnetofonových nahrávek z let 1994–1999 zachycujících 294 mluvčích (rodilých Brňanů). Rozsah korpusu je zhruba 600 tisíc pozic. Všechny texty byly digitalizovány i ve zvukové podobě, která je k dispozici v ÚČJ FF MU.

BMK byl ve snaze o kompatibilitu vytvářen v souladu s hlavními zásadami už dříve vytvořeného *Pražského mluveného korpusu* (PMK), přístupného rovněž v rámci ČNK. Znamená to především, že se snažil ve vyvážených proporcích obsáhnout čtyři sociolingvistické proměnné: pohlaví mluvčího, věk (ve 4 věkových kategoriích), vzdělání (ve 3 kategoriích) a 2 typy promluvy: formální a neformální. BMK obsahuje 135 tzv. formálních nahrávek (vytvářených sledem odpovědí na otázky kladené podle jednotného dotazníku) a 115 nahrávek neformálních (tvořených tematicky volným dialogem blízkých osob). Každá nahrávka BMK byla dále doplněna zpřesňujícími informacemi o mluvčích, o roku svého vzniku, případně relevantními údaji o situaci promluvy (tyto informace jsou v ČNK skryté).

Pravidla přepisu nahrávek v BMK také v základních rysech odpovídala pravidlům užívaným v PMK, šlo tedy o účelovou kombinaci fonetického zápisu a standardních pravopisných norem (detailnější informace o přepisu v BMK je možno vyhledat na výše zmíněné internetové adrese). Specifika BMK spočívají kromě

drobností především v pokusu o nahrazení tradiční interpunkce interpunkcí pauzovou a ve striktním zachycování simultánnosti dialogických promluv. V řešení obou jevů mají jak pravidla PMK, tak pravidla BMK své přednosti i nevýhody. Způsob zvolený v PMK do jisté míry znásilňuje skutečnost (interpunkce podle pravopisné normy je do značné míry umělá, zvláště v neformálních dialogických promluvách nevystihuje reálně členění mluveného jazyka; nezdá se také pravděpodobné, že by se ve spontánním dialogu neobjevila simultánnost, kterou PMK nezachycuje), na druhé straně způsob užitý v BMK se ve výsledku ukázal jako ne příliš vhodný pro korpusové zpracování jazykového materiálu. Pauzová interpunkce je náročnější, což vedlo (i přes několik sjednocujících kontrol) k diferencím zápisu u jednotlivých přepisovatelů. Co se týče simultánnosti, v podstatě jakýkoli grafický systém jejího zachycení, který je naprosto transparentní v souvislých přepisech celého dialogu, je většinou méně transparentní pro uživatele korpusu pracujícího s konkordančními seznamy. Způsob zápisu, který byl zvolen v BMK, sice s korpusovým užitím počítal a je v každém dokladu jednoznačně dešifrovatelný, poněkud však komplikuje vyhledávání reálných konkordancí (v případě, kdy je kontinuita textu narušena zařazením simultánního úseku jiné repliky) a vyžaduje od uživatele větší soustředěnost při filtraci získaných dat. Nabízí se tedy otázka k diskusi, zda regulovaná rezignace na přesnost zachycení jazykového projevu nemůže být při přípravě relativně rozsáhlých korpusových materiálů někdy přípustná, ba dokonce výhodná, jestliže zjednodušuje práci jak tvůrcům, tak i uživatelům korpusu a není v rozporu s hlavním účelem, který má korpus plnit (většina korpusů je zatím utvářena hlavně jako materiálový zdroj pro studium jevů lexikálních a morfologických).

BMK je v ČNK zatím uložen bez morfologického značkování. Užití morfologického analyzátoru vytvořeného pro spisovný jazyk je v případě BMK komplikováno nejen obecně mluvenostními rysy textů, ale také zvláště výraznou hláskovou, tvarovou i lexikální variabilitou brněnské mluvy (tj. prolínáním dialektických, interdialektických, obecněčeských i spisovných podob). Na Fakultě informatiky Masarykovy univerzity se už delší dobu připravují úpravy analyzátoru *ajka* (vytvořeného primárně pro morfologickou analýzu spisovného jazyka; viz např. Sedláček – Smrž, 2001), které by se měly vyrovnat alespoň s některými specifiky brněnské mluvy a umožnit v co největším rozsahu její automatickou analýzu (detailněji např. Hlaváčková, 2002). I v případě užití tohoto analyzátoru bude však nutné automatickou analýzu doplnit ručním značkováním.

2. Korpus soukromé korespondence

Soukromá korespondence přináší řadu informací o jazykovém úzu, které lze jen obtížně čerpat z jiných zdrojů. K jejím přednostem patří především značná autenticita, neformálnost a spontánnost projevu (v tomto smyslu jsou se soukromými dopisy srovnatelné snad jen tajně pořízené nahrávky běžné mluvy). Významná je též pestrost teritoriální příslušnosti pisatelů, která umožňuje alespoň doplňkové studium územně diferenčních jazykových jevů, aniž by bylo nutno

provádět náročný terénní výzkum. Lingvisticky relevantním znakem soukromé korespondence je i její oscilace mezi mluveností a psaností, nespisovností a spisovností. Korespondenční texty mají značnou přitažlivost také pro klasickou stylistiku, textovou lingvistiku, pragmalingvistiku apod. Kromě zmíněných hodnot je vytváření korpusu soukromé korespondence v současné době cenné i tím, že možná v poslední fázi zachycuje tradiční ručně psané dopisy a zároveň mapuje první stadium korespondence využívající elektronická média.

2.1 Sběr materiálu

Korpus soukromé korespondence vzniká v ÚČJ FF MU od konce 90. let 20. století. V současné době se opírá o archiv obsahující zhruba 3000 elektronických přepisů ručně psaných dopisů, dále 1500 e-mailů a 1000 SMS zpráv. Datace shromážděné korespondence se pohybuje v rozmezí posledních 15 let, autoři pocházejí z celého území České republiky, jsou to většinou mladí lidé a převažují mezi nimi ženy. Dopisy jsou sbírány anonymně, jejich dárci uvádějí na kartičky připojené ke každému dopisu pouze standardizované základní údaje o pisateli a adresátovi.

Legislativní (i morální) problém se zveřejněním soukromé korespondence je řešen tím, že dopisy jsou získávány od adresátů, tedy se svolením alespoň jednoho účastníka komunikace. Sám adresát z nich navíc vyškrtá všechny identifikační údaje (pokud to neudělá důsledně, jsou eliminovány ve fázi přepisu dopisu).

2.2 Přepis

Při přepisu do elektronické verze je striktně dodržována původní podoba dopisů (pouze identifikační údaje jsou nahrazeny sjednocujícím znakem a skrytou vyvolatelnou informací, zda jde o příjmení, adresu, telefonní číslo apod.; nedořešenou otázkou zůstává, mají-li být ponechány, nebo odstraněny přezdívky). Zvláštními znaky jsou zachycovány také informace o grafické úpravě dopisů (např. o textovém členění pomocí odstavců, o typech písma, o obrázcích). Zvlášť označovány jsou citátové pasáže (v rozsahu věty a více), které budou v korpusu sice odhalitelné, ale nebudou podléhat běžnému vyhledávání a statistickým analýzám, aby nezkreslovaly obraz jazyka soukromé korespondence.

Přepisy důsledně zachovávají i pravopisné chyby. Tento postup je samozřejmě diskutabilní: na jedné straně je jistě informativní – umožňuje porovnávat dodržování pravopisných norem v klasické a e-mailové korespondenci nebo u různých věkových a vzdělanostních vrstev pisatelů, pomáhá hledat současné pravopisné tendence, což lze využít pro kodifikační účely nebo pro školskou praxi, apod. Na druhé straně pravopisné chyby komplikují automatickou analýzu textu při morfologickém značkování (buď analyzátoru vůbec znemožňují rozpoznat špatně napsané slovo – např. *nábitek*, což je ještě ta lepší varianta, protože pak na takové slovo upozorní právě fakt, že zůstalo neoznačkováno, nebo, což je horší možnost, vedou analyzátor k mylným závěrům – např. grafická podoba *noví* je analyzována jako nominativ plurálu životných maskulin, ale v dopise

může jít o chybný zápis spisovného nominativu, příp. u neživotných i akuzativu singuláru maskulin, nebo o chybný zápis celé řady obecněčeských tvarů; takové mylně označkované případy se pak v korpusu obtížně odhalují). Pravopisné chyby mohou způsobovat komplikace i uživatelům elektronických korpusů, protože problematizují vyhledání a usouvztažnění všech výskytů téhož slova nebo tvaru při frekvenčních analýzách. Možným a asi nejvhodnějším řešením by bylo vytvářet hned při přepisu do elektronické podoby v případech narušení pravopisné normy dvě propojené podoby – pravopisně správnou a reálnou. V přípravě brněnského KSK se však postupovalo trochu jinak, tj. (ne)dodržování pravopisných pravidel se zachycovalo ve shodě s realitou a řešení pravopisné substandardnosti bylo přesunuto až do fáze morfologického značkování, resp. disambiguace. V této fázi je pak k pravopisně chybné formě, pokud je ovšem odhalena, přidáváno pravopisně správné lemma a zvláštní poznámka informující o pravopisné nekorektnosti. Lemma pak umožní při vyhledávání usouvztažnit graficky chybnou podobu s podobami správnými.

Poměrně složitá pravidla jsou užívána pro přepis e-mailů nedodržujících diakritiku českých slov (v e-mailech, které vůbec neužily diakritiku, je diakritika doplňována, v e-mailech, které ji užily částečně, a je tedy zřejmé, že technické vybavení její užití umožňovalo, je ponechávána reálný stav. Je to samozřejmě opět problematické řešení mající svá pro a proti podobně jako zmíněné zachycování pravopisných chyb). SMS zprávy jsou zpracovávány zcela samostatným způsobem do zvláštní databáze, která nebude součástí elektronického korpusu (v SMS zprávách totiž ještě výrazněji narůstá problém s diakritikou, připojuje se komplikace časté nestandardní zkratkovistosti a především splývavého zápisu, který znemožňuje delimitaci slov – viz např. klasické kreativní spojení *jaxemáš*).

2.3 Základní parametry připravovaného korpusu korespondence

Ze shromážděného přepsaného materiálu v současné době vzniká korpus 2000 klasických dopisů a 1000 e-mailů celkově reprezentujících 3 000 různých pisatelů. (Zajištění diferenčnosti pisatelů při anonymním sběru korespondence vyžadovalo mnohdy náročnou ruční kontrolu dopisů vykazujících shodu v obecných sociolingvistických charakteristikách, o nichž viz dále.) Tento korpus obsahuje v části shromažďující klasické dopisy cca 940 tisíc pozic, v části e-mailů cca 220 tisíc pozic. Zastoupeny jsou v něm všechny věkové kategorie, akcentována je však korespondence mladých lidí, která nejlépe dokládá vývojové tendence češtiny a také nejlépe vypovídá o proměnách žánru a stylu v souvislosti s přechodem mezi klasickou korespondencí a korespondencí elektronickou. Část korpusu obsahující elektronický přepis klasických dopisů by měla být v budoucnu propojena s digitálními fotokopiemi originálů, které už jsou k tomuto účelu připraveny.

2.4 Sociolingvistické charakteristiky

Každý dopis v korpusu je označen kombinací značek reflektujících sociolingvistické parametry. Tvoří je: pohlaví, věk (4 kategorie) a vzdělání (3 kategorie)

pisatele i adresáta (v těchto informacích korpus udržuje kompatibilitu s PMK a BMK), dále teritoriální (nářeční) zázemí pisatele (zpracované do číselných kódů podle nářečních oblastí v *Českém jazykovém atlase*), typ vztahu mezi pisatelem a adresátem a rok napsání dopisu. Vyhledávací program *Bonito – Manatee* (autor Pavel Rychlý, viz např. Rychlý, 2000; Rychlý – Smrž 2004), pod nímž je pracovní verze korpusu uložena, umožňuje v případě potřeby pracovat pouze s vybranou částí textů na základě zadané kombinace těchto charakteristik. Stávající velikost korpusu pochopitelně nedovoluje smysluplně využít všechny kombinace parametrů, už teď se ale ukazuje jako statisticky relevantní např. vyhledávání teritoriálně nebo genderově podmíněných jazykových jevů.

2.5 Morfologické značkování

V současné době se začíná pracovat také na morfologickém značkování korpusu korespondence, konkrétně té části, která obsahuje ručně psané dopisy. Automatická analýza je ztěžována hraničním postavením korespondenčních textů mezi psaností a mluveností. To s sebou přináší v ještě větší míře nežli u mluveného jazyka střídání standardního (spisovného) jazykového kódu s kódy substandardními (často i v rámci jediné věty, např. *...učitelé jsou někteří dobrý...*). Při analýze korpusu zahrnujícího dopisy pisatelů z Čech, Moravy a Slezska je nutno počítat nejen s prolínáním spisovné a obecné češtiny, ale i s potenciálním výskytem lexikálních, morfologických a hláskových prvků všech nářečí. Dopisy mladých lidí navíc často obsahují kreativní okazionalismy, citátová cizojazyčná slova, atypicky adaptované přejímky apod. V neposlední řadě ztěžují automatické zpracování i zatemňující pravopisné chyby, o nichž už byla řeč.

Morfologické značkování korpusu korespondence je zatím v pokusné fázi. Podobně jako při značkování BMK se užívá modifikovaná verze automatického morfologického analyzátoru *ajka* (detailněji Hlaváčková – Sedláček, 2004).

Analyzátor *ajka* se opírá o morfologickou databázi *i_par* (autor Marek Veber), která vychází z algoritmického popisu české formální morfologie (Osolsobě, 1996). Podstatou je segmentace slov na kmen a koncovku. Kmen je ještě dále členěn na neměnnou základní část a na tu část, která se během flexe mění (nejčastěji kvůli hláskovým alternacím), tzv. intersegment. Výsledkem segmentace jsou tři inventáře segmentů – koncovkové množiny, intersegmenty a vlastní kmenové základy. *I_par* obsahuje slovník kmenů (opírající se o SSJČ), s nímž jsou propojeny množiny intersegmentů a koncovek a pravidla (vzory) určující, které z možných kombinací intersegmentů a koncovek jsou přípustné pro daný kmenový základ.

Při automatické analýze textu je pak každá slovní forma analyzátorem zkoumána odzadu, a pokud je identifikována, je k ní přiřazeno lemma, tj. základní tvar (nominativ, infinitiv) a dále značky pro gramatickou charakteristiku zkoumaného tvaru. Protože jsou slovní formy zkoumány bezkontextově, v silně homonymní češtině přiřadí analyzátor ke zkoumanému slovnímu tvaru zpravidla více interpretací. Kontextové zjednoznačnění – tzv. desambiguace – se pak provádí buď ručně, nebo zčásti automaticky, např. na základě využití syntaktických pravidel.

V první fázi byl korpus soukromé korespondence označkován neupraveným analyzátorem určeným pro spisovnou češtinu. Po vytřídění slov, která zůstala zcela bez morfologické značky, byl vytvořen jejich frekvenční seznam. Při přípravě modifikovaného analyzátoru se pak pracovalo jen s těmi, která měla frekvenci minimálně 5. V seznamu neoznačkovaných forem se objevila jednak slova lišící se od spisovné češtiny pouze koncovkou, jednak slova lišící se jiným způsobem, tj. hláskovou změnou kmene, slovotvornou či lexikální diferencí (slova pouze pravopisně substandardní byla vytříděna zvlášť). Úpravy analyzátoru reflektují uvedené dvě základní skupiny trochu rozdílně: případy nespisovných koncovek připojovaných ke spisovnému kmeni řeší rozšířením koncovkových množin analyzátoru o varianty substandardních útvarů, ostatní případy zahrnutím do databáze *i_par*. V prvním případě analyzátor k morfologickým značkám automaticky přidává atribut upozorňující na koncovkovou nespisovnost, ve druhém případě bude připojena skrytá, ale vyvolatelná poznámka o substandardnosti. Tyto informace by měly při vyhledávání v korpusu umožňovat identifikaci slovních forem neodpovídajících spisovné češtině. Přitom je třeba říci, že všechna textová slova v korpusu, která se liší od spisovné varianty jen pravidelnou nářeční či obecněčeskou koncovkou nebo pravidelnou nářeční hláskovou obměnou kmene, příp. pouze pravopisem, by měla mít v konečné fázi spisovné lemma (narozdíl od nářečních, slangových a různě okazionálních slovotvorných a lexikálních variant).

Z koncovek nespisovných útvarů byly do analyzátoru v první fázi přidány zatím především některé koncovky obecné češtiny a středomoravských dialektů, tj. kódů, které se na podobě jazyka soukromé korespondence vedle spisovné češtiny podílejí nejvíce (jednak proto, že dopisy z území, kde jsou tyto kódy základem běžné mluvy, jsou v korpusu zastoupeny nejpočetněji, jednak proto, že obecná čeština a zčásti i středomoravský dialekt mají větší tendenci projevit se v psané korespondenci nežli další územně podmíněné variety; o tom viz Hladká – Šindlerová, 2004).

Po zatím provedených úpravách je morfologický analyzátor schopen rozpoznat o cca 40 tisíc slovních výskytů více nežli před úpravami. Je však třeba si uvědomit, že rozšířením „rozsahu" analyzátoru na druhé straně vzrůstá počet slov nabízených k desambiguaci.

Dosud provedené změny analyzátoru zdaleka neřeší automatické značkování korpusu korespondence v komplexnosti. Analyzátor bude dále upravován (i v souvislosti s úpravami pro korpusy mluveného jazyka), impulsy k vylepšení jistě přinesou i zkušenosti s desambiguací, s níž se teprve začíná. Bez zapojení ruční práce se však značkování korpusu korespondence podobně jako značkování korpusů běžné mluvy neobejde.

***

Nejbližším cílem zpracovávání korespondenčních textů v Ústavu českého jazyka na Filozofické fakultě Masarykovy univerzity v Brně je do konce roku 2005 (v rámci projektu GAČR 405/03/0248 *Současná soukromá korespondence. Vytvoření databáze a zpracování vybraných jevů z pohledu lexikologicko-lexikografického a dialektologického*, s jehož podporou je vypracována i tato stať)

připravit a pokud možno i morfologicky označkovat korpus klasické korespondence a předat ho do ČNK. Dále uložit korpus klasických dopisů společně s korpusem e-mailů, sbírkou SMS zpráv a databází dalších údajů týkajících se soukromé korespondence na CD, které by mělo být k dispozici odborné veřejnosti.

Druhá část přednášky, o niž se opírá tento příspěvek, byla věnována konkrétním příkladům využití korpusu soukromé korespondence pro lingvistický výzkum. Z prostorových důvodů je zde neuvádíme a odkazujeme na články, které se této problematice věnují detailněji. Všechny dosavadní analýzy byly zatím prováděny pouze na cvičných sondách, tj. na souborech majících rozsah maximálně 500 dopisů. Ukázaly např.: 1) využitelnost korpusu korespondence pro lexikologicko-lexikografické účely, konkrétně pro mapování expresivní a kolokviální vrstvy slovní zásoby, kterou tradiční česká lexikografie – mj. pro nedostatek vhodných referenčních zdrojů – dosud poněkud opomíjela (např. Hladká, 2000; Hladká 2005); 2) přínosnost korpusu korespondence pro poznávání vztahu dichotomií psanost – mluvenost a spisovnost – nespisovnost a pro odhalování teritoriálních diferencí ve funkční škále užívání spisovné češtiny a substandardních útvarů národního jazyka (např. Hladká, 2001; Hladká – Šindlerová, 2004); 3) možnosti využití korespondenčních textů pro studium některých pragmatických aspektů komunikace, např. pro hledání genderových diferencí v komunikačních strategiích (např. Hladká, 2004). Soukromá korespondence nabízí využití ještě v mnoha dalších směrech. Relevantnost dat zjistitelných z elektronických korpusů epistolárních textů však bude limitována rozsahem a kvalitou těchto korpusů. V tomto smyslu je výše popsaný brněnský pokus pouze iniciačním krokem na možné cestě.

LITERATURA

HLADKÁ, Zdeňka: *Několik poznámek k výběru lexikálních jednotek pro výkladové slovníky.* In: O. Martincová – J. Světlá (eds.), Nová slovní zásoba ve výkladových slovnících. Praha : ÚJČ AV ČR 2000, s. 35–42.

HLADKÁ, Zdeňka.: *Spisovnost a nespisovnost v jazyce soukromé korespondence (se zřetelem k teritoriální příslušnosti pisatelů).* Naše řeč, 5, 84, 2001, s. 225–234.

HLADKÁ, Zdeňka: *Korpus soukromé korespondence jako zdroj poznání jazykového úzu.* In: M. Šimková (ed.), Tradícia a perspektívy gramatického výskumu na Slovensku. Bratislava : Veda 2003, s. 130–135.

HLADKÁ, Zdeňka: *Soukromá korespondence z hlediska rodových diferencí.* In: V. Patráš (ed.), Súčasná jazyková komunikácia v interdisciplinárnych súvislostiach. Banská Bystrica : Univerzita Mateja Bela 2004, s. 469–475.

HLADKÁ, Zdeňka: *Univerbizace – korpusy – slovníky (malá materiálová sonda).* In: R. Blatná – V. Petkevič (eds.), Jazyky a jazykověda. Sborník k 65. narozeninám prof. Františka Čermáka. Praha : ÚČNK FF UK 2005, s. 503–514.

HLADKÁ, Zdeňka – ŠINDLEROVÁ, Hana: *Jakou češtinou si dopisujeme na Moravě.* In: J. Fiala (ed.), AUPO, Fac. Phil., Moravica 1. Olomouc : UP 2004, s. 105–114.

HLAVÁČKOVÁ, Dana: *Korpus mluvené češtiny.* Diplomová práce na FF MU (rkp.), Brno 1998.

HLAVÁČKOVÁ, Dana: *Morfologické značkování korpusu brněnské mluvy.* In: Z. Hladká – P. Karlík (eds.), Čeština – univerzália a specifika, 4, Praha : Nakladatelství Lidové noviny 2002, s. 311–312.

HLAVÁČKOVÁ, Dana – SEDLÁČEK, Radek: *Morfologické značkování korpusu soukromé korespondence.* Příspěvek přednesený na XIV. kolokviu mladých jazykovedcov v Šintavě u Seredi, Slovensko – 8.–10. 12. 2004. V tisku.

OSOLSOBĚ, Klára: *Algoritmický popis české formální morfologie a strojový slovník češtiny.* Disertační práce na FF MU (rkp.), Brno 1996.

PETKEVIČ, Vladimír: *Perspektivy morfologického značkování (českých korpusů).* Příspěvek přednesený na pracovním semináři „Obecné a specifické aspekty tvorby korpusů českého jazyka". Praha 25. 3. 2004.

RYCHLÝ, Pavel: *Korpusové manažery a jejich efektivní implementace.* Disertační práce na FI MU (rkp), Brno 2000.

RYCHLÝ, Pavel – SMRŽ, Pavel: *Manatee, Bonito and Word Sketches for Czech.* In Proceedings of the Second International Conference on Corpus Linguisitcs. Saint-Petersburg : Saint-Petersburg State University Press 2004, s. 124–132.

SEDLÁČEK, Radek – SMRŽ, Pavel: *A New Czech Morphological Analyser ajka.* In: Proceedings of of the 4th International Conference TSD 2001, Berlin : Springer Verlag 2001, s.100–107.

CREATING CORPUSES OF CZECH AT THE DEPARTMENT OF CZECH LANGUAGE, FACULTY OF ARTS, MASARYK UNIVERSITY BRNO

The contribution deals with two electronic corpuses of Czech created at the Faculty of Arts, Masaryk University Brno, namely the corpus of Brno urban speech and the corpus of private correspondence. The contribution deals with problems of collecting the data, transferring them into an electronic form and their morphological tagging. Possibilities of further use of both corpuses, especially the corpus of private correspondence, are also indicated.

Zdeňka Hladká
Ústav českého jazyka FF MU
Arna Nováka 1
602 00 Brno