Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Závěrečná zpráva o realizaci projektu[1]

1. Stručný přehled splněných cílů projektu Centra komputační lingvistiky

a. přehled splněných cílů (v souladu s návrhem projektu a uzavřenou smlouvou, časový postup)

Cílem Centra komputační lingvistiky byl výzkum a vývoj v oblasti moderní počítačové lingvistiky na zcela nově získané úrovni založené na jedinečné vícerovinné analýze velmi rozsáhlého korpusu. Činnost Centra měla a má velký význam pro aplikace v mnoha oborech služeb a průmyslu, které pracují s komunikací člověka s počítačem. Hlavními tématy výzkumu byly následující cíle; jak ukázala veřejná odborná rozprava uskutečněná ve dnech 29. a 30. listopadu 2004 za účasti 9 zahraničních odborníků (zápis přiložen), všechny cíle byly beze zbytku splněny, a to na vysoké odborné úrovni.

(A) Teoretické aspekty komputační lingvistiky se zaměřením především na češtinu v podobě psané i mluvené a s ohledem na možné aplikace. Tento výzkum bylo možno vést na kvalitativně vyšší úrovni, než kdy bylo možné, díky existenci Pražského závislostního korpusu (PDT), který nabízí možnost poloautomatické analýzy velkého souboru stovek tisíc českých vět; PDT byl vytvořen právě v Centru (viz bod (B) níže).

Teoretický výzkum v rámci Centra byl neoddělitelně spjat s řešenými projekty, a to jednak jako předpoklad pro jejich formulaci a teoretický základ pro jejich řešení, jednak tyto projekty přinášeli vedle ověřování platnosti navržených hypotéz i důležité další podněty pro teoretické bádání a pro obohacení daného pojmového rámce.

(B) Stěžejním projektem CKL bylo vybudování Pražského závislostního korpusu (Prague Dependency Treebank, PDT) – morfologicky (1 974 301 slov / 116 885 vět) a analyticky (1 507 372 slov / 87 898 vět) anotovaný korpus PDT byl obohacen o značkování na tektogramatické rovině, která zachycuje význam vět (49 192 vět), a to včetně koreferenčních vztahů v textu a aktuálního členění. Tato mimořádně rozsáhlá počítačová korpusová data jsou jedinečná zvláště proto, že čeština je jediný jazyk s bohatou morfologií, který byl analyzován v podobné míře. PDT byl vytvořen ve spolupráci s Ústavem formální a aplikované lingvistiky (ÚFAL) na Matematicko-fyzikální fakultě UK pro účely podrobné gramatické, sémantické a lexikální analýzy češtiny (CD-ROM s korpusem PDT patří k zásadním publikovaným výsledkům CKL, viz Příloha 1.).

(C) Metodologie výzkumu v rámci Centra byla založena na prohloubeném studiu, porovnávání a kvalifikovaném využití postupů strukturních i statistických včetně metod strojového učení, s ohledem na specifické typologické vlastnosti češtiny jako vysoce flexívního jazyka. V tomto ohledu byla vyvinuta originální metodologie, neboť dosud známé přístupy čerpaly ze zdrojů v angličtině a jiných jazycích s nízkým stupněm flexe (především s "chudou" morfologií a pevným pořádkem slov).

(D) Potřebná pozornost byla věnována matematickým a komputačním základům metod a algoritmů komputační lingvistiky a postupů zpracování přirozeného jazyka. Byl vyvinut nástroj pro systém strojového učení na základě tzv. průměrovaného perceptronu, a dále byly testovány na dalších jazycích nástroje založené na exponenciálním statistickém modelu. Byl rovněž prováděn výzkum v oblasti využití konečných automatů a převodníků pro jazykové modelování a morfologické značkování.

[1] Zpráva podepsaná řešitelem, která byla schválena oponentním řízením, se současně se zápisem z oponentního řízení, vyúčtováním za uplynulé období, se zasílá písemně i elektronicky zadavateli.

(E) Jedním ze základních směrů Centra byl statisticky založený výzkum v oblasti rozeznávání mluvené řeči, který také patří k výsledkům Centra majícím největší význam pro spojení úsilí odborníků z této oblasti a badatelů v lingvistice a informatice.

Zvláštní pozornost byla věnována studiu tzv. suprasegmentálních jevů, jako je význam větné prozodie (pro který je neobyčejně výhodným východiskem pražská analýza aktuálního členění věty (na ‚danou' a ‚novou' informaci, tj. na základ a ohnisko) a studiu modelování jazyka, opět se zřetelem na flexívní povahu češtiny.

Zásadním přínosem bylo zapojení Centra do mimořádně rozsáhlého mezinárodního projektu MALACH (Multilingual Access to Large Spoken Archives), jehož cílem je vývoj systémů pro automatický přepis svědeckých výpovědí lidí, kteří přežili holocaust. Svědecké výpovědi byly pořízeny ve více než 30 různých jazycích a česká strana je prostřednictvím Centra spoluzodpovědná za zpracování jazyků střední a východní Evropy.

(F) Dalším cílem výzkumu Centra bylo vytváření a využívání vícejazyčných zdrojů. Pozornost byla věnována zejména studiu a uplatnění paralelních korpusů se zaměřením na strojový a strojem podporovaný překlad (CD-ROM věnovaný nástrojům pro strojový překlad viz Příloha 1.; projekt překladu mezi blízkými jazyky Česílko) a dalším aplikacím jako vyhledávání informací (data mining) ve vícejazyčných textech. Takto pojatá výzkumná činnost vedla k získání dalších znalostí o češtině srovnatelných s výsledky výzkumu jiných jazyků

(G) Činnost Centra bohatě naplnila očekávané možnosti v navazování a udržování těsných kontaktů s českým a mezinárodním průmyslem využívajícím počítače, o čemž svědčí i zájem partnerů a uživatelů z oblasti aplikační sféry o vhodně zpracované a užitečné zdroje pro široce založený vývoj a aplikace.

Časový postup řešení stanovených cílů bylo v zásadě dodrženo, drobné úpravy vyplývaly z úrovně dosažených výsledků a byly vždy specifikovány v upřesnění dílčích cílů projektu pro následující rok řešení projektu.

b. přehled nesplněných úkolů

Všechny úkoly specifikované v návrhu projektu a v uzavřené smlouvě byly splněny. Pokud byly v průběhu programu přijaty změny oproti zadání, vždy šlo o rozšíření činnosti, nikoli o redukci či změnu úkolů (viz bod 1.e níže).

c. zhodnocení výsledků a plnění cílů projektu

CKL se stalo unikátním výzkumným pracovištěm v ČR v oblasti komputační lingvistiky a automatického zpracování přirozeného jazyka, a to především tím, že (i) v něm byl integrován a vzájemně posilován výzkum jak jazyka psaného, tak mluveného, (ii) že výzkum má pevné a originální teoretické základy v oblasti gramatiky i lexika, ale neztrácí ze zřetele i aspekty aplikační, a (iii) že jsou ve vzájemné rovnováze jak lingvistické, tak i informatické aspekty tohoto výzkumu. V tomto smyslu se dá říci, že v této integraci má CKL i významné postavení mezi předními světovými centry výzkumu v komputační lingvistice, o čemž svědčí i zájem o výsledky výzkumu v CKL (viz bod 4. níže).

Dlouhá léta se vyvíjely metody zpracování řeči a jazyka odděleně a cíle obou proudů byly odlišné. Jestliže metody zpracování řeči v oblasti automatického rozpoznávání usilovaly o co nejnižší chybovost funkce klasifikátoru a obvykle nevyužívaly podpory lingvistických metod (morfologické, syntaktické, sémantické analýzy...), pak při zpracování jazyka se obvykle

pracovalo s textem, který byl gramaticky správně zapsán, a řešila se úloha syntaktické a sémantické analýzy poměrně snadno vymezitelné věty.

Vybudováním CKL na pracovištích MFF UK a ZČU, které mají dlouhodobé zkušenosti i prokazatelné výsledky ve zpracováním jednak jazyka a jednak mluvené řeči, bylo dosaženo velmi účinného napojení obou řešitelských týmů. Pracovníci CKL tak mohli okamžitě využívat know-how jak z oblasti zpracování řeči, tak zpracování jazyka, které byla na těchto pracovištích aktuálně k dispozici. To značně zrychlilo a zefektivnilo navazující výzkumné práce.

Zázemí obou výzkumných pracovišť (MFF UK a ZČU), šířka výzkumného záběru a prokazatelné, světově srovnatelné výsledky jednoznačně ukazují, že Centrum komputační lingvistiky vyrostlo na pevných základech jak Pražské lingvistické školy na MFF UK, tak mnohem mladší, ale neméně úspěšné školy řečových technologií na ZČU v Plzni. Podobná spojení lingvistických a řečových výzkumných týmů lze pozorovat v poslední době i na mnoha zahraničních univerzitách. Z tohoto hlediska jde o průkopnické a v ČR jedinečné a zcela nezastupitelné pracoviště.

O hodnotě výsledků získaných v rámci Centra bezesporu svědčí i vysoká míra zájmu o spolupráci ze strany zahraničních i českých výzkumných pracovišť, o zapojení CKL do mezinárodních projektů (podrobněji viz bod 4.), i zájem o využití konkrétních výsledků projektu (viz bod 1.d.)

V CKL bylo za dobu jeho trvání připraveno celkem 259 publikací, (r. 2001 – 41, r. 2002 – 68, r.2003 – 72, r.2004 - 78). Seznam konkrétních výsledků a výstupů získaných v rámci realizace projektu (včetně kódů výsledku v registru RIV) jsou uvedeny v přílohách – Příloha 1. Seznam dat a nástrojů získaných v rámci realizace projektu a Příloha 2. Seznam publikací.

d. konkrétní využití dosažených výsledků a výstupů projektu

Uvádíme zde seznam nejdůležitějších výstupů Centra s krátkým popisem a strukturou uživatelů (více viz bod 2.d. způsob využívání výsledků a výstupů projektu aplikační sférou a v rámci regionu a zejména Příloha 1. Seznam dat a nástrojů získaných v rámci realizace projektu).

Pražský závislostní korpus, verze 1.0 (PDT 1.0), publikováno v roce 2001, vydáno LDC: byly podepsány licence o výzkumném využití se 112 uživateli, z toho 31 uživatelů v ČR a SR, 36 v Evropě, 30 v Americe, 13 v Asii, 3 na Středním východě. Součástí PDT 1.0 je rovněž řada nástrojů na zpracování češtiny, vč. morfologického slovníku a analyzátoru (více viz Příloha 1.).

Prague Czech English Dependency Treebank (PCEDT), paralelní, česko-anglický závislostní korpus, který byl v listopadu 2004 vydán v Linguistic Data Consortium (LDC, LDC2004T25, ISBN: 1-58563-321-6). Bezprostředně po vydání přišlo 5 objednávek na PCEDT (Kanada, 2× USA, Hong Kong, Nizozemsko).

Valenční slovník českého jazyka VALLEX, verze 1.0: VALLEX 1.0 byl publikován na sklonku roku 2003 a od té doby se zaregistrovalo přes 80 uživatelů (domácích i zahraničních pracovišť) (více viz Příloha 1.).

Vyhledávací nástroj NetGraph: NetGraph server má doposud 25 registrovaných uživatelů a byl instalován v ÚJČ a v LDC, Philadelphia, USA. NetGraph klient je instalovaný i na FF UK, v JÚĽŠ AV SR a na Univerzitě Komenského v Bratislavě. V současné době je NetGraph používán pro češtinu, slovenštinu a arabštinu; plánuje se jeho použití v rámci prohledávání kompletního Českého národního korpusu. LDC uvažuje o jeho použití pro další jazyky (více viz Příloha 1.).

Grafický anotovací nástroj TrEd používají kromě dalších pracovišť UK (např. ÚTKL FF UK) kolegové v Josef Stefan Institute, Ljubljana, Slovinsko; JÚĽŠ, Bratislava, Slovensko; Sárská

univerzita v Saarbrückenu, Německo; od těchto uživatelů existuje rovněž zpětná vazba z hlediska vývoje tohoto nástroje (více viz Příloha 1.).

Nástroj pro zpracování XML souborů XSH má řádově stovky uživatelů; více o nástroji viz Příloha 1.).

Nástroj pro vytváření anotovaných korpusů ACT byl instalován ve Slovanském ústavu AV.

Czech Broadcast News Speech, Czech Broadcast News Transcripts: korpus řečových nahrávek pro přípravu systémů rozpoznávání řeči, vydáno LDC v r. 2004.

e. změny proti zadání v realizaci projektu provedené v období řešení projektu

V roce 2001 byla při oponentním řízení schválena Oponentní radou korekce plánu v oblasti rozpoznávání řeči – CKL se od roku 2002 zapojilo do vysoce prestižního projektu MALACH, jehož cílem je vývoj systémů pro automatický přepis svědeckých výpovědí lidí, kteří přežili holocaust. Svědecké výpovědi byly pořízeny ve více než 30 různých jazycích a česká strana je spoluzodpovědná za zpracování jazyků střední a východní Evropy. Na projektu participují Visual History Foundation v Hollywoodu, Johns Hopkins University v Baltimore, University of Maryland, IBM, MFF UK v Praze a ZČU v Plzni. Anotační práce na zpracování svědeckých výpovědí jsou podporovány National Science Foundation (USA), Project #0122466.

CKL bylo pověřeno organizací **Mezinárodního kongresu lingvistů**, CILXVII, který se konal 24.-29.7.2003 pod patronátem mezinárodní organizace Comité International Permanent des Linguistes. Kongresu se zúčastnilo 436 účastníků, k jeho konání byl vydán sborník abstraktů (ed. Hajičová, CKL, 2003, tištěná forma) a sborník příspěvků (eds. Mírovský, Kotěšovcová, Hajičová, CKL, 2003, CD-ROM).

V souladu s cíli CKL (výzkum a vývoj v oblasti počítačové lingvistiky s důrazem na budování jazykových korpusů a souvisejících nástrojů a jejich další využití) došlo v roce 2003 k rozšíření činnosti CKL: CKL se zapojilo do spolupráce s Ústavem formální a aplikované lingvistiky MFF UK a Ústavem srovnávací jazykovědy FF UK, jejímž cílem je budování Pražského závislostního korpusu pro arabštinu, **Prague Arabic Dependency Treebank**.

2. Personální a organizační zabezpečení činnosti centra (aktuální stav v porovnání s výchozími podmínkami na začátku sledovaného období)

a. vývoj ve složení pracovního týmu z hlediska kvalifikace ve vztahu k pracovní náplni v centru a pracovnímu zařazení

Po celou dobu realizace projektu bylo personální zabezpečení Centra stabilní, vyvážené a vzhledem k pracovní náplni adekvátní.

	celkový počet pracovníků	celkový počet úvazků	pracovníci s úvazkem ≥ 0,7	pracovníci s úvazkem < 0,7
2000	38	20,75	18	20
2001	42	23,25	25	17
2002	41	29,45	29	14
2003	40 (+21)	24,7 (+10,75)	24	37
2004	50	29,65	24	26

Celková výše úvazků se v době řešení projektu Centra podle plánu mírně zvyšovala (při zachování rozpočtu v oblasti mezd), v roce 2003 byli z důvodů zabezpečení organizace Mezinárodního kongresu lingvistů někteří dlouhodobě spolupracující studenti zaměstnáni na větší část roku na částečné úvazky (čísla v závorkách). Plánovaný nárůst počtu úvazků byl z velké části využit pro zaměstnání dlouhodobě spolupracujících studentů po ukončení magisterského studia.

Odbornou úroveň CKL po celou dobu jeho trvání zajišťovali čtyři profesoři podílející se na jeho úkolech a dalších pět pracovníků s vědeckou hodností CSc, resp. Dr. či Ph.D. V době trvání Centra se dva z jeho pracovníků habilitovali (2003, doc. Hajič, doc. Štícha) a pět jeho pracovníků obhájili disertační práce s tématikou bezprostředně související s úkoly Centra (2001 Kuboň, Lopatková, 2004 Ircing, Ondruška, Ribarov); další disertační práce byla odevzdána v říjnu 2004 (Zeman, obhajoba začátkem roku 2005). Řada dalších členů pracovního týmu v současné době dokončuje své disertační práce.

b. vývoj ve složení týmu z hlediska věku

Pracovní tým CKL svým složením odpovídal podmínce zaměstnávat především mladé vědecké a odborné pracovníky. Většina týmu se rekrutovala ze studentů a doktorandů MFF UK, FF UK a ZČU, pro něž bylo zaměstnání v CKL jejich prvním zaměstnáním. Následující tabulka tento trend potvrzuje – v průběhu celé doby řešení projektu byli přijímáni noví mladí pracovníci z řad dlouhodobě spolupracujících studentů, kteří dokončili magisterské studium.

	počet pracovníků mladších než 35 let	počet úvazků	%
2000	32	16,7	84%
2001	31	17,2	74%

2002	25	22,8	75%
2003	25 (+21)	17,1 (+10,75)	69%
2004	39	21,4	72%

Další objem práce odváděli studenti magisterského studia (OON), pro něž možnost zapojit se do projektů CKL představovala významný impuls pro orientaci na další vědeckou práci.

c. změny v řídící a organizační struktuře centra a jejich přínos či nedostatky

Organizační struktura Centra i jeho řízení se ukázalo jako plně funkční, proto zůstalo stabilní a neměnné po celou dobu trvání projektu.

d. vytvořená nová pracovní místa

Na dobu realizace Centra bylo vytvořeno 30 plných úvazků, na kterých bylo zaměstnáno celkem 50 pracovníků (včetně 3 pracovníků, kteří zajišťovali technickou podporu CKL). Dalších 20-24 pracovníků (v naprosté většině studenti magisterského, případně doktorandského studia, kteří dlouhodobě spolupracovali na úkolech Centra) mělo uzavřeno smlouvu o pracovní činnosti či provedení práce. (Číselné údaje jsou za rok 2004.)

Pokud se nepodaří získat finanční prostředky z programu MŠMT Centra typu A., všechna pracovní místa na MFF UK a na FAV CZ zaniknou s koncem projektu Centra komputační lingvistiky. Na pracovišti UJČ se podařilo vzniklé částečné úvazky částečně převést na úvazky ústavní, a tudíž budou zachovány.

3. Přístrojové vybavení a technické zabezpečení činnosti centra

Pracoviště MFF UK Praha:

Před zahájením řešení projektu v roce 2000 bylo pracoviště na MFF vybaveno odpovídajícím způsobem, umožňujícím veškeré potřebné aktivity díky jiným projektům – například výzkumnému záměru, realizovanému v rámci UK v Praze, MFF. Výpočetní síla dostupných serverů byla spíše na dolní hranici potřebných kapacit, nicméně každoročně bylo možné provést částečnou obnovu a upgrade.

V průběhu realizace plánu výzkumného centra bylo pracoviště dovybaveno pro řešení rozsáhlejších problémů, vyžadujících výkonnější stroje a větší množství pracovníků. Byl vybudován centrální datový prostor s dostatečným zálohováním a kapacitou, v důsledku vývoje bezpečnostní situace v počítačové síti Internet bylo také řešeno zabezpečení proti virům i aktivním útokům hackerů.

V průběhu celých pěti let byla veškerá zakoupená technika maximálně využívána a investiční plány byly dynamicky upravovány podle aktuálních potřeb a cen. Predikovat technologický vývoj v horizontu pěti let není snadné, nicméně celkové odhady potřebných prostředků byly správné. Příslušné změny v původních plánech jsou zdokumentovány ve výročních zprávách CKL.

V posledním roce řešení projektu byly změny jen drobné. Prostředky, určené na zvýšení kapacity datového serveru byly částečně přesunuty na nákup výpočetních serverů, jelikož bylo nalezeno levnější řešení, než jsme předpokládali. Obdobně byly použity i prostředky, původně plánované na upgrade síťové technologie, jelikož ten byl realizován ze zdrojů Univerzity Karlovy v rámci rekonstrukce budovy MFF UK.

Detailní specifikace zakoupeného vybavení (investice 2004, 2 900 tis. Kč):
- Jolly NAS 6,4 TB (diskové pole)
- 3x výpočetní server (2x CPU Opteron, 64-bit, 16GB RAM), částečně hrazeno i z dalších projektů
- 1x výpočetní server HP (4x CPU Opteron, 64-bit, 32 GB RAM)
- 4x notebook
- 3x nové PC (upgrade pracovní stanice na řadu P4 nemožný)
- 4x upgrade pracovní stanice

Nyní, na konci projektu lze říci, že pracoviště disponuje dostatečným vybavením a je schopno se podílet na všech mezinárodních výzkumných aktivitách v dané oblasti. Jsou k dispozici i potřebné prostředky pro výuku a prezentaci. Vzhledem k rychlosti vývoje informatiky však není možné ustrnout v současném stavu – bez náležité údržby a obnovy by až doposud vynaložené prostředky mohly být ztraceny. Předpokládáme, že projekty, navazující na práci vykonanou v rámci CKL, umožní také příslušnou údržbu a další rozvoj.

Pracoviště ZČU Plzeň:

Základní vybavení investičními prostředky bylo získáno při startu Centra a v každém roce byly tyto prostředky doplňovány a inovovány. Vzhledem k tomu, že pracovníci Centra komputační lingvistiky participovali na řešení rozsáhlého projektu „MALACH“, kde byly řešeny úlohy akustického a jazykového modelování, a to vedle češtiny i několika dalších evropských jazyků, byl pro tyto účely zapotřebí značný výpočetní výkon. Navíc, na řešení úloh Centra se podílel stále

větší počet studentů magisterského a doktorského studia, kteří plně využívali sice starší, ale stále plně funkční investiční prostředky zakoupené v minulém období.

Detailní specifikace zakoupeného vybavení (investice 2004, 400 tis.Kč):

V roce 2004 byly na účet spoluřešitele (ZČU) převedeny hlavním řešitelem (MFF UK) investiční prostředky dle původního plánu ve výši 400 tis. Kč. Za zmíněné investiční prostředky byla provedena inovace výpočetních stanic CKL:

- 2 x Pracovní stanice FSC Scenic W620 pro práci s velkými korpusy (Pentium4-3.4HT, 2GB RAM, 2x 160GB, SATA disk, DVD?RW(DL), Acer 19" LCD, WindowsXP Pro MUI)
- Pracovní stanice AMD pro práci s velkými korpusy (AMD Athlon 64 FX 53 (2,4 GHz 64bit), 4GB RAM, 2x 160GB SATA disk, DVD?RW(DL), Acer 19" LCD, WindowsXP Pro CZ)
- Výpočetní server AMD pro kompozici rozsáhlých konečných automatů (AMD Athlon 64 3500+ (2,2 GHz), 2GB RAM, 200GB SATA disk + 120 GB U-ATA, DVD?RW(DL), Acer 19" LCD, WindowsXP Pro CZ)
- Výpočetní server P4 pro kompozici rozsáhlých konečných automatů (Pentium4-3.2HT, 2GB RAM, 200GB SATA disk, DVD?RW(DL), Acer 19" LCD, WindowsXP Pro CZ)

Pracoviště UJČ Praha:

Oddělní jazykové kultury a oddělení gramatiky byly z financí centra postupně dovybaveny několika počítači, klasickými i LCD monitory, laserovými tiskárnami, cestovní tiskárnou, několika notebooky a množstvím drobné výpočetní techniky (USB paměti aj.).

Notebooky (s jednou cestovní přenosnou tiskárnou) byly zakoupeny zejména pro nové pracovníky jazykové poradny a pro dva mladé lingvisty z nově vzniklého oddělení gramatiky, pro potřeby spolupráce s externími pracovníky v rámci úkolů řešených v CKL a pro potřeby ukládání dílčích korpusů. Zejména pro práci s nově vznikajícími počítačovými korpusy textů byly zakoupeny přídatné velkokapacitní paměti.

Specifikace zakoupeného vybavení (investice 2004, 200 tis.Kč):

Z financí určených na investice byl objednán dataprojektor pro nově zřízené oddělení gramatiky, notebook pro nového pracovníka oddělení gramatiky a počítač pro dr. Uhlířovou, jímž bude nahrazen její dosavadní počítač starý.

Využití přístrojového vybavení po ukončení řešení projektu

Získané přístrojové vybavení bude po skončení projektu Centra využíváno v aktivitách realizovaných v rámci výzkumných záměrů Informatické sekce MFF IU a Katedry kybernetiky FAV ZČU, v případných dalších aktivitách, o jejichž podpoře se nyní rozhoduje (výzkumné centrum typu A, granty GAČR a granty GA UK, projekty Informatické společnosti) a při zpracování diplomových a disertačních prací studentů MFF UK a FAV ZČU.

4. ***Spolupráce centra*** (prokažte zájem partnerů a uživatelů Vašich výsledků konkrétními údaji o naplňování jejich požadavků *:*

a. úroveň odborné spolupráce v rámci ČR, s ostatními zakládajícími a spolupracujícími organizacemi ve sledovaném období

Spolupráce Centra se zakládajícími organizacemi (Univerzita Karlova, Západočeská univerzita, Ústav pro jazyk český AV Praha) byla bezproblémová a užitečná pro obě strany. Zakládající organizace přispívaly dotacemi podle původního plánu (místnosti, sdílení počítačového vybavení, využívání služeb centrálních oddělení fakult, rektorátu či Ústavu jako jsou knihovny, účetnictví apod.), na druhé straně Centrum poskytovalo řadu příležitostí pro studenty příslušných vysokých škol pracovníky Ústavu nezačleněné do Centra, a to jak vedením výuky, diplomových prací a projektů (i na jiných fakultách), tak i poskytováním příležitostí účasti na mezinárodních akcích Centra. Je třeba rovněž ocenit, že UK a její matematicko-fyzikální fakulta přispívala k výměně zahraničních expertů (z Fondu mobility rektora UK byl např. hrazen jeden z dlouhodobých pobytů).

Probíhala také spolupráce s univerzitami a AV ČR na projektech základního výzkumu a dále na využití Českého národního korpusu a prostředků pro jeho analýzu a pro značkování textů, zaměřené k využití při počítačovém zpracování češtiny, mj. pro strojový překlad, pro vyhledávání informací, "data mining" a pro komunikaci s inteligentními databázemi. Vyměňovaly se výsledky výzkumu Centra a Ústavu českého národního korpusu FF UK i Ústavu teoretické a komputační lingvistiky FF UK (vzájemně), výsledky Centra slouží i Ústavu pro jazyk český AV ČR (včetně výuky a školení jeho mladých pracovníků).

b. nová zapojení do mezinárodních struktur ve sledovaném období

V době realizace projektu CKL navázalo spolupráci s následujícími univerzitami a výzkumnými pracovišti mimo území ČR:

(A) Dohody o spolupráci:

- **JUĽŠ SAV**, Bratislava a pedagogická fakulta Univerzity Komenského, Bratislava v rámci programu vědecko-technické spolupráce se Slovenskem (KONTAKT/Slovensko 2004-2005), výměna expertů
- **Projekt MALACH** – Multilingual Access to Large Spoken Archive, NSF USA: Visual History Foundation, CA, USA; IBM Research, NY, USA; Johns Hopkins University, MD, USA; University of Maryland, MD, USA (2001-2006)
- **Projekt Strojový překlad**, KONTAKT/NSF (MŠMT), Johns Hopkins University, MD, USA, 2003-2005
- Dohoda o publikaci lingvistických dat s **Linguistic Data Consortium**, Philadelphia, PA, USA (1x CD-ROM vydáno 2001, 4x CD-ROM vydáno 2004, dále bude vydáno 1x CD-ROM v r. 2005)

(B) Spolupráce s univerzitami a výzkumnými pracovišti (včetně hlavních bodů spolupráce):

- Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA (prof. Frederick Jelinek) – zpracování mluvené řeči, projekt MALACH, strojový překlad, generování v přirozeném jazyce;
- University of Maryland, College Park, MD, USA (prof. Doug Oard) – projekt Malach a strojový překlad pro účely vyhledávání informací;

- IBM Research, Yorktown Heights, NY, USA (Michael Picheny) – projekt MALACH, metodika transkripce mluveného textu;
- Visual History Foundation Survivors of the Shoa, North Hollywood, CA, USA (Sam Gustman) – zdroje spontánní mluvené řeči ve velkém rozsahu;
- Ohio University, Bloomington, OH, USA (Jiří Hana) – vývoj morfologického analyzátoru ruštiny;
- Istituto di Linguistica Computazionale, C.N.R., Pisa, Italy (prof. Antonio Zampolli a prof. Nicoletta Calzolari) – slovníky, korpusová lingvistika, anotování korpusů;
- Natürlichsprachliche Systeme, Fachbereich Informatik, Universität Hamburg, Germany (prof. Walter von Hahn, prof. W. Menzel) – komputační lingvistika, automatická syntaktická analýza, formální gramatiky;
- Computational Linguistics and Phonetics, Universität des Saarlandes, Saarbrücken, Germany (prof. Hans Uszkoreit, prof. M. Pinkal) – formální sémantika, komputační lingvistika; anotování korpusů na hloubkové úrovni;
- Institut für Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken, Germany (prof. H. Gerzymisch-Arbogast) – překladové systémy;
- Institut National des Langues et Civilisations Orientales (INALCO), Paris, France (prof. Patrice Pognan) – komputační lingvistika, práce s PDT;
- Xerox Research Center Europe, Language Research Group, Grenoble, France (prof. Jean-Pierre Chanod) – metody založené na konečně-stavových automatech, strojový překlad;
- Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, U.S.A. (prof. Martha Palmer) – komputační lingvistika, tree adjoining grammars, proposition bank;
- Linguistic Data Consortium, University of Pennsylvania, Philadelphia, U.S.A. – anotování korpusů, distribuce lingvistických dat;
- Linguistic Department of the University of Massachusetts, Amherst, USA (prof. Barbara Partee) – formální sémantika;
- School of Informatics, University of Edinburgh, Great Britain (prof. Mark Steedman) – komputační lingvistika;
- Department of Linguistics, University of Uppsala, Sweden (prof. Anna Sågvall Hein) – komputační lingvistika, slovníky;
- Jazykovedný ústav Ľudovíta Štúra, Akadémia vied Slovenskej republiky, Univerzita Komenského, Bratislava, Slovenská republika (Dr. Mária Šimková, Ing. Vladimír Benko) – korpusová lingvistika, anotování slovenských dat;
- CKL je členem mezinárodní sítě ENABLER (European National Activities for Basic Language Resources), která si klade za cíl zintenzívnit spolupráci mezi národními centry vyvíjející a zpracovávající jazykové zdroje.

CKL bylo delegováno MŠMT jako reprezentant v celoevropském projektu programu ERA (LangNet, koordinace a evaluace projektů Language Technology).

Vedle výše uvedených institucionálních kontaktů mají pracovníci CKL aktivní osobní pracovní kontakty s vědci a pedagogy dalších předních světových univerzit, jako jsou např.

- v USA: Stanfordova universita, Columbia University, Harvard University, Massachusetts Institute of Technology, University of California San Diego, Johns Hopkins University
- v Japonsku: Kyoto University
- v Jižní Koreji: Seoul National University
- v Austrálii: Centre for Language Technology, Macquarie University
- v Německu: univ. v Mnichově, v Bonnu, v Heidelbergu, v Lipsku, v Postupimi, v Bochumi, Humboldtova univ.
- ve Francii: univ. v Grenoblu
- v Itálii: univerzita v Benátkách, Univerzita v Pise
- v Maďarsku: budapeštská univerzita, univerzita v Szegedu, Morphology Inc. (Budapešť)
- v Polsku: varšavská univerzita, univerzita v Krakově
- v Rusku: Moskevská státní univerzita

- ve Velké Británii: University College, Londýn, univerzita v Cambridge, v Edinburghu, v Lancasteru, v Manchesteru, v Brightonu, v Sheffieldu
- v Bulharsku: Bulharská akademie věd
- v Slovinsku: Univerzita v Lublani
- v Dánsku: Kodaňská univerzita

c. kvalita spolupráce s aplikační sférou a v rámci regionu

K využití výsledků Centra došlo především ve spolupráci s výzkumnými pracovišti v zahraničí (viz výše o využití výsledků na pracovištích na Slovensku, ve Slovinsku, ve Spojených státech, ve Francii, v Německu); na výsledky CKL (především na Pražský závislostní korpus, PDT) navazují další výzkumné projekty. Mladí pracovníci vyškolení v Centru uplatňují získané know-how již dnes v komerční sféře (např. ve výzkumném středisku IBM v Praze). Vzhledem k tomu, že se v CKL řeší problematika základního výzkumu, poskytujeme zatím výsledky pro výzkumné účely bezúplatně.

K využití know-how došlo například i při přípravě rozsáhlého korpusu řečových nahrávek pro přípravu systémů rozpoznávání řeči, viz bod 1.d). Tento korpus se prodává v LDC (University of Pennsylvania, Philadelphia, USA). Podpora CKL při přípravě tohoto korpusu je v LDC uvedena.

CKL představuje výzkum, jehož se účastní i mladí pracovníci, kteří následně získané know-how využívají ve výzkumných a vývojových odděleních komerčních firem, které se zabývají různými úlohami zpracování přirozeného jazyka v mluvené i psané podobě.

d. způsob využívání výsledků a výstupů projektu aplikační sférou a v rámci regionu

Během realizace CKL byla vyvinuta řada nástrojů, dat a postupů, které jsou využívány na lingvistických pracovištích v České republice i v zahraničí. Nejdůležitější výstupy CKL s charakteristikou uživatelů jsou uvedeny v bodu 1.d, seznam výsledků viz Příloha 1., Seznam dat a nástrojů získaných v rámci realizace projektu.

Dále byla navázána cenná spolupráce s aplikační sférou, a to v rámci republiky i v měřítku mezinárodním. V následující tabulce uvádíme seznam subjektů, které mají zájem o využití výsledků Centra, včetně oblasti zájmu.

Název	Sídlo	Oblast zájmu
Skřivánek, s.r.o.	Na dolinách 22, 14700 Praha 4	Automatizace překladu
LEDA, s.r.o.	26301 Voznice 64	Lingvistická podpora elektronických slovníků
NetCentrum s.r.o.	Drtinova 10 15000 Praha 5	Vyhledávání v textech na Centrum.cz
ASPI Publishing, s.r.o.	U nákladového nádraží 6 Praha 3	Právní informační systémy: jazyková podpora a rozpoznávání řeči
LANGMaster International, s.r.o.	Branická 107 14700 Praha 4	Automatický překlad, podpora výukového software
SpeechTech, s.r.o.	Morseova 5 30100 Plzeň	Rozpoznávání řeči
IBM ČR, s.r.o.	V parku 4 14200 Praha 4	Porozumění mluvené řeči

ARTLingua	Myslíkova 6, 12000 Praha 2	Strojový překlad
Microsoft Corp.	Dublin, Irsko	Testovací jazykové korpusy
Reader's Digest Výběr, s.r.o.	V celnici 4, 11000 Praha 1	Překlad produkce do jazyků střední a východní Evropy, překlad z angličtiny do češtiny
Visual History Foundation	100 Universal City Plaza, Bldg. 5225 Room 149 Universal City, CA 91608	Rozpoznávání řeči pro vyhledávání v rozsáhlých audioarchívech v češtině a jazycích střední Evropy

5. ***Podpora mladých výzkumných pracovníků*** (aktuální stav v porovnání s výchozími podmínkami na začátku sledovaného období)

a. doktorské studijní programy

Dvě sekce CKL při univerzitách (MFF UK a ZČU) byly po celou dobu realizace projektu významným způsobem zapojeny do programů Doktorského studia. Podíleli se na výchově mladých výzkumných pracovníků, jimž jednak poskytovali příležitost k vlastnímu bádání a k jeho prezentaci, jednak umožňovali jejich zapojení do větších výzkumných úkolů.

Řada pracovníků Centra se podílela na výuce v rámci magisterských i doktorských studijních programů – na MFF UK jde o magisterský a doktorský obor Matematická lingvistika, na FAV ZČU o magisterský program Kybernetika a řídicí technika a doktorsky program Kybernetika. Například v letním semestru 2003/2004 vedli celkem 28 přednášek a seminářů na MFF UK, FF UK a KK ZČU, v zimním semestru 2004/2005 vedli 23 přednášek a seminářů na MFF UK, FF UK a KK ZČU. Mimo to byli vedoucími řady studentských projektů. Díky zapojení pracovníků Centra do výuky se podstatným způsobem (o 12 přednášek/seminářů oproti roku 2000) rozšířila nabídka přednášek a seminářů pro studenty magisterského i doktorského studia se zájmem o komputační lingvistiku.

CKL po celou dobu svého trvání podporovalo zapojování studentů magisterského studia do doktorských studijních programů. Témata jejich disertačních prací úzce souvisela s vědeckým programem Centra, jejich školiteli byli pracovníci CKL. Např. v roce 2003 pracovníci Centra školili 28 doktorandů, 5 z nich obhájilo své disertační práce (další práce je odevzdaná). V roce 2004 to bylo 21 interních a 13 externích doktorandů.

Pro jednotlivé školitele-pracovníky Centra uvádíme v závorce počet doktorandů: prof. Hajičová (10), doc. Hajič (16), dr. Kuboň (3), dr. Lopatková (1), prof. Panevová (5), prof. Psutka (6), dr. Vidová Hladká (2).

b. podíl mladých výzkumníků (do 35 let), vč. objemu prací a pracovní kapacity, způsob podpory jejich odborné práce ze strany centra.

Podpora mladých pracovníků byla jednou z priorit Centra. Jak bylo uvedeno již v bodu 2.b., pracovní tým CKL se z více jak 70% (přepočteno podle velikosti úvazků) skládal z mladých výzkumníků do 35 let. Většina pracovníků se rekrutovala ze studentů a doktorandů MFF UK, FF UK a ZČU, pro něž bylo zaměstnání v CKL jejich prvním zaměstnáním. Další objem práce odváděli studenti magisterského studia (OON), pro něž možnost zapojit se do projektů CKL představovala významný impuls pro orientaci na další vědeckou práci.

Lze tedy říci, že podpora mladých pracovníků Centra byla velmi intenzívní. Vedle pravidelných seminářů, na nichž vystupovali se svými referáty, podíleli se mladí pracovníci podstatnou měrou i na výjezdních seminářích pracoviště. Jejich výzkum je integrální součástí vědeckých úkolů Centra, jde o práci navýsost týmovou, takže jsou v denním pracovním kontaktu se svými vedoucími i dalšími klíčovými pracovníky projektu.

O velmi dobrých výsledcích výzkumné práce mladých pracovníků i o jejím ohlasu na mezinárodním poli svědčil i počet přijatých referátů na mezinárodních konferencích; účast mladých pracovníků na těchto konferencích byla velmi hojná a aktivní, což bylo umožněno finanční podporou z prostředků Centra (viz bod c).

c. Podpora mladých výzkumných pracovníků (konkrétní příklady ve sledovaném období)

(A) Účast studentů a doktorandů na mezinárodních konferencích a workshopech v zahraničí

V průběžných zprávách byl uveden vždy výčet zahraničních cest mladých vědeckých pracovníků, které Centrum umožnilo (přehled pro rok 2004 je uveden v oddíle specifikace a zdůvodnění jednotlivých výdajových položek ve vztahu k projektu, formulář F3C-čerpání), zde uvádíme souhrn pro jednotlivé pracovníky:

Emanuel Beška – 1x konference/workshop v zahraničí
Alena Böhmová – 4x konference/workshop v zahraničí
Ondřej Bojar – 1x konference/workshop v zahraničí
Markéta Ceplová – 1x konference/workshop v zahraničí
Ondřej Cikhart – 1x konference/workshop v zahraničí
Martin Čmejrek – 9x konference/workshop v zahraničí
Jan Cuřín – 8x konference/workshop v zahraničí
Jiří Hana – 1x konference/workshop v zahraničí
Jiří Havelka – 3x konference/workshop v zahraničí
Martin Holub – 3x konference/workshop v zahraničí
Petr Homola – 4x konference/workshop v zahraničí
Pavel Ircing – 1x konference/workshop v zahraničí
Kolář – 1x konference/workshop v zahraničí
Ivona Kučerová – 2x konference/workshop v zahraničí
Lucie Kučová – 1x konference/workshop v zahraničí
Kateřina Marková – 1x konference/workshop v zahraničí
J. Mlejnecká – 1x konference/workshop v zahraničí
Roman Ondruška – 3x konference/workshop v zahraničí
Petr Pajas – 2x konference/workshop v zahraničí
Pavel Pecina – 1x konference/workshop v zahraničí
Nino Peterek – 1x konference/workshop v zahraničí
Petr Podveský – 2x konference/workshop v zahraničí
M. Pravdová – 1x konference/workshop v zahraničí
M. Prošek – 2x konference/workshop v zahraničí
Veronika Kolářová (Řezníčková) – 6x konference/workshop v zahraničí
Kiril Ribarov – 4x konference/workshop v zahraničí
Karolina Skwarska – 1x konference/workshop v zahraničí
Jiří Semecký – 1x konference/workshop v zahraničí
Kamila Smejkalová – 2x konference/workshop v zahraničí
Otakar Smrž – 5x konference/workshop v zahraničí
Pavel Straňák – 1x konference/workshop v zahraničí
Markéta Straňáková – 3x konference/workshop v zahraničí
Jan Štěpánek – 2x konference/workshop v zahraničí
Kateřina Veselá – 2x konference/workshop v zahraničí
Dan Zeman – 4x konference/workshop v zahraničí
Zdeněk Žabokrtský – 4x konference/workshop v zahraničí

(B) Pracovní pobyty studentů a doktorandů v zahraničí

Petr Biskup – 1x pracovní pobyt v zahraničí
Martin Čmejrek – 2x pracovní pobyt v zahraničí
Jan Cuřín – 1x pracovní pobyt v zahraničí
Eva Flanderková – 2x pracovní pobyt v zahraničí
Martin Holub – 3x pracovní pobyt v zahraničí
Petr Homola – 2x pracovní pobyt v zahraničí
Pavel Ircing – 1x pracovní pobyt v zahraničí
Ivona Kučerová – 2x pracovní pobyt v zahraničí
Pavel Květoň – 1x pracovní pobyt v zahraničí
Pavel Machek – 1x pracovní pobyt v zahraničí
J. Mlejnecká – 1x pracovní pobyt v zahraničí
Jiří Mírovský – 1x pracovní pobyt v zahraničí
Roman Ondruška – 1x pracovní pobyt v zahraničí
Petr Pajas – 1x pracovní pobyt v zahraničí
Josef V. Psutka (jun.) – 1x pracovní pobyt v zahraničí
Veronika Kolářová (Řezníčková) – 3x pracovní pobyt v zahraničí
Kamila Smejkalová – 1x pracovní pobyt v zahraničí
Otakar Smrž – 1x pracovní pobyt v zahraničí

Pavel Straňák – 1x pracovní pobyt v zahraničí
Barbory Vidová-Hladká – roční postdoc pobyt na Johns Hopkins University
Zdeněk Žabokrtský – 6x pracovní pobyt v zahraničí

(C) Účast studentů a doktorandů na mezinárodních letních školách v zahraničí:

Ondřej Bojar – 3x letní škola
Martin Čmejrek – 1x letní škola
Jiří Havelka – 4x letní škola
Martin Holub – 1x letní škola
Petr Homola – 3x letní škola
Václav Honetschlager – 1x letní škola
Jiří Kocanda – 1x letní škola
Pavel Machek – 1x letní škola
Petr Podveský – 2x letní škola
Veronika Kolářová (Řezníčková) – 1x letní škola
Otakar Smrž – 1x letní škola
Jan Štěpánek – 2x letní škola

(D) Publikace doktorandů, společné publikace doktorandů se školiteli:

Centrum podporovalo publikační činnost mladých spolupracovníků, ať už samostatnou, nebo se školiteli, čímž podstatným způsobem přispívalo k jejich odbornému růstu. V seznamu publikací za dobu trvání Centra je 192 položek, jejichž autory či spoluautory jsou doktorandi a mladí vědečtí pracovníci.

(E) Podpora nových projektů podávaných mladými pracovníky:

CKL podporovalo mladé spolupracovníky a doktorandy při řešení nových projektů souvisejících s jejich odbornými zájmy. K těmto projektům patřil zejména:

- Projekt Prague Arabic Dependency Treebank, viz bod 1.e zprávy a zpráva za rok 2003
- Projekt ACT, viz zpráva za rok 2003

Dále Centrum podporovalo podávání nových grantů, jejichž řešiteli jsou doktorandi účastnící se práce CKL (zejména GAUK - 1 přijatý projekt, 8 nových podaných projektů).

6. **Způsoby zpřístupnění výsledků a výstupů centra veřejnosti** (aktuální stav v porovnání s výchozími podmínkami na začátku sledovaného období – konkrétní akce pro odbornou i laickou veřejnost, internet. adresy ...)

(A) Mezinárodní konference a workshopy:

Pracovníci CKL se zúčastnili řady mezinárodních konferencí a jiných odborných setkání, na kterých přednesli zvané přednášky (např. v roce 2003 12 zvaných přednášek, v roce 2004 14 zvaných přednášek) a recenzované příspěvky o výsledcích dosažených v projektech CKL, případně prezentovali své výsledky na posterech. Participovali také na workshopech při konferencích, které byly vynikající příležitostí pro sdílení nových výsledků a postupů. Z nejprestižnějších konferencí oboru jmenujme následující:

- konference pořádané **Association for Computational Linguistics** (ACL, NAACL, EACL)
- konference pořádaná **International Speech Communication Association** (ISCA)
- **Coling** (konference pořádaná International Committee of Computational Linguistics, ICCL)
- **MT Summit** (pořádaný Association for Machine Translation, AMTA/EAMT)
- **Language resources and Evavuation Conference** (pořádá ACL/ELRA)
- **Text, Speech, Dialogue** (pořádané FI MU/ZČU)
- **Světový kongres lingvistů**

(Seznamy konkrétních pracovních cest pracovníků CKL jsou uvedeny ve zprávách za jednotlivé rok projektu, včetně titulů prezentovaných příspěvků; cesty za rok 2004 jsou uvedeny ve formuláři F3C-čerpání, bod 2. specifikace a zdůvodnění jednotlivých výdajových položek ve vztahu k projektu).

(B) Publikace:

Publikace v domácích i zahraničních časopisech a ve sbornících mezinárodních konferencí zpřístupňují výsledky Centra široké odborné veřejnosti. Za dobu trvání Centra bylo připraveno 259 publikací (seznam viz Přílohu 2).

(C) Technické zprávy:

CKL vydávalo technické zprávy (ve spolupráci s ÚFALem MFF UK) o dílčích výsledcích výzkumu; Za dobu existence Centra bylo vydáno 17 technických zpráv, které jsou k dispozici jednak tištěné, jednak na adrese http://ckl.mff.cuni.cz:8080/pub/publications.jsp?type=tr. Citace technických zpráv jsou uvedeny v seznamu publikací v Příloze 2.

(D) Webové stránky CKL:

Byly vytvořeny www stránek CKL, http://ckl.mff.cuni.cz/, které podávají komplexní informaci o činnosti Centra. Kromě základních informací o struktuře, výzkumných tématech a cílech CKL zde lze nalézt stručný přehled dílčích cílů projektu, a potom zejména odkazy na stránky jednotlivých úkolů řešených v rámci CKL (PDT, ČAK, VALLEX, PADT, MALACH), kde jsou také k dispozici volně šiřitelné nástroje (morfologická analýza, taggery, editory stromových struktur – TrEd, Graph, internetový prohlížeč stromů Netgraph). K dispozici je rovněž nově implementovaná databáze publikací pracovníků CKL s elektronickými verzemi jednotlivých příspěvků (pokud to umožňuje nakladatel). K nahlédnutí jsou i průběžné oponentní zprávy Centra.

(E) ENABLER Network

CKL jako člen mezinárodní sítě ENABLER Network (zaměřené na shromažďování informací o existujících jazykových zdrojích pro jednotlivé jazyky) průběžně aktualizovalo přehled dostupných jazykových zdrojů, které vznikaly v rámci jednotlivých projektů Centra, čímž zásadně přispívalo k

informovanosti o svých výsledcích a výstupech. Poskytlo k dalšímu využití zejména následující výstupy:

- Pražský závislostní korpus (PDT)
  - teoretické základy
  - vlastní data
- paralelní data angličtina – čeština
- slovníky
  - valenční slovník užívaný při anotování PDT-VALLEX
  - komplexně anotovaný valenční slovník VALLEX
- nástroje pro anotaci korpusu, vyhledávání v korpusech

(F) Podobný význam bude mít i účast CKL, resp. v navazujícím Ústavu formální a aplikované lingvistiky v programu ERA (LangNet).

(G) Den otevřených dveří:

CKL každoročně v rámci Dnu otevřených dveří na MFF UK širší odborné veřejnosti živou formou představovala svou činnost. Zájemci především ze středních škol byli seznamováni s tématy, na nichž se v Centru pracuje, a také s výsledky Centra.

7. Závěrečné zhodnocení

a. programu Výzkumná centra LN jako celku, jeho celkový přínos vědecké sféře

Program Výzkumná centra LN jako celek byl při svém vyhlášení i průběhu významným příspěvkem k rozvoji vědeckého výzkumu v České republice: poskytl především do té doby neprůchodnou možnost vytvořit pracovní místa pro mladé vědecké pracovníky a finančně je zabezpečit na slušné úrovni. Jsme přesvědčeni, že pro mnohé z nich byl tento program pobídkou, aby svou vědeckou erudicí, schopnostmi i zápalem přispěli k zaplnění generační mezery, kterou v mnoha oborech – především humanitního charakteru – u nás zanechalo totalitní období. Přitom program dal těmto mladým lidem možnost být v co nejtěsnějším kontaktu s kolegy v zahraničí, představovat svou práci na významných zahraničních konferencích, a tak získávat další podněty pro práci vlastní. Bylo velmi důležité, že Výzkumná centra mohla vznikat jak se zaměřením na teoretický výzkum, tak i v oblasti výzkumu aplikovaného. Organizačně i finančně tento program podpořil spolupráci pracovišť univerzitních i akademických, v projektech aplikovaného výzkumu pak i spolupráci se sférou aplikační.

V závěru čtyřapůlletého období, po které program probíhal, se však projevila řada nevysvětlitelných problémů, ukazujících na zřejmá (snad administrativní) opomenutí, která se v posledním roce nepříznivě podepsala na průběhu práce existujících Center. Bylo od počátku jasně dáno, že Centra byla vytvořena na určitý časový úsek; jejich pracovníci však – mimo jiné i na základě zkušeností ze zahraničí – důvodně očekávali, že práce Center bude důkladně prověřena, zahraničními oponenty srovnána s výsledky v zahraničí a po tomto zhodnocení bude úspěšným Centrům umožněno formulovat návazný program, v němž by uplatnění našli jak již zaškolení mladí pracovníci Center dosavadních, tak také další mladí adepti vědy. V jistém smyslu nás v tomto přesvědčení utvrzoval i velmi obsáhlý dotazník MŠMT, který jsme vyplňovali v lednu 2004, a který při pečlivém a pravdivém vyplnění přinesl poskytovateli dotace jistě dobrý přehled o tom, jak bylo dotace využito (bohužel jsme se však o dalším osudu dotazníku či o jeho využití nic nedověděli).

Kolem vyhlášení návazného programu bylo však stále mnoho nejasností, oddalování a protichůdných informací, které vyústily ve vyhlášení jen samostatného programu aplikovaného výzkumu (přitom se téměř ve stejném termínu sešly návrhy na Centra i na výzkumné záměry, což určitě k dobrému a účelnému rozvrhování témat i personálního obsazení neprospělo). Teprve v průběhu letních prázdnin, s krajně napjatým termínem podávání návrhů byl vyhlášen program center typu A, který je svou koncepcí ovšem původnímu záměru Center jako středisek mladé vědy s určitou vyhraněnou koncepcí a za vedení zkušených vědeckých kapacit velmi vzdálen; svými podmínkami připomíná v podstatě soubor doktorandských a postdoktorandských grantů. Nejvíc zarážející je skutečnost, že MŠMT nepočítá se zahraniční oponenturou předložených návrhů (ani abstrakt návrhu neměl být podán v cizím jazyce, což svědčí o uspěchanosti a nedomyšlenosti celého programu, která snadno povede k oprávněným námitkám, jak vůbec mohly být vybrány nejlepší projekty ve srovnání se zahraničím).

b. činnosti vlastního centra výzkumu z hlediska zhodnocení jeho přínosu za celé období řešení projektu

Pokud jde o vlastní činnost Centra komputační lingvistiky, lze bez nadsázky konstatovat, že se podařilo shromáždit odborně velmi vyspělou početnou skupinu mladých nadějných vědeckých pracovníků, kteří v průběhu projektu prokázali svou odbornou erudici, svou schopnost samostatné vědecké práce a v neposlední řadě své nadšení pro týmovou spolupráci nad perspektivními a ve světě s vynikajícím ohlasem přijímanými projekty. Byl vytvořen u nás jediný integrovaný tým pro výzkum psané i mluvené řeči. Jak ukázala veřejná vědecká rozprava o výsledcích Centra konaná ve dnech 29.-30.listopadu 2004, za účasti 7 předních zahraničních vědců z oboru komputační lingvistiky, tyto výsledky mají přední místo v evropském i světovém

výzkumu a bylo by třeba, aby mladí pracovníci Centra i jejich pokračovatelé dostali možnost v tomto výzkumu pokračovat.

Příloha 1. Seznam dat a nástrojů získaných v rámci realizace projektu,

- **Pražský závislostní korpus, PDT 1.0** (PDT, http://ufal.mff.cuni.cz/pdt)
  RIV/00216208:11320/01:00105063
  Pražský závislostní korpus, PDT 1.0 vydalo LDC v roce 2001 (katalogové číslo LDC2001T10, ISBN: 1-58563-212-0) obsahuje
  -- data:
  - anotovaná data: texty anotované na morfologické (1 974 301 slov / 116 885 vět) a analytické (1 507 372 slov / 87 898 vět) rovině, ukázka anotací na tektogramatické rovině
  - neanotované texty
  - česko anglický paralelní korpus

  -- nástroje
  - NetGraph (vyhledávání na stromech)
  - Tred (stromový editor, vyhledávání na stromech)
  - morfologický analyzátor
  - taggery (zjednoznačnění morfologické informace)

  -- dokumentace

- **Pražský závislostní korpus, PDT 2.0**
  RIV/zatím nepřiděleno
  Pražský závislostní korpus, verze 2.0 je stěžejním výsledkem práce Centra. Jde o obohacení korpusu PDT, verze 1.0 o anotaci na tektogramatické rovině. PDT 2.0 bude vydáno v LDC v roce 2005. PDT 2.0 obsahuje
  -- data:
  - texty anotované na tektogramatické rovině (49 192 vět)

  -- nástroje
  - nové, podstatně rozšířené verze nástrojů NetGraph (viz níž), Tred (viz níž), morfologický analyzátor, taggery

  -- dokumentace

- **Prague Arabic Dependency Treebank, PADT 1.0,**
  RIV/zatím nepřiděleno
  http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T23
  Závislostní korpus moderní standardní arabštiny vzniká s využitím bohatých zkušeností a nástrojů získaných při vytváření PDT ve spolupráci s Ústavem srovnávací jazykovědy FF UK a Linguistic Data Consorcium. Korpus je morfologicky anotován pomocí nástroje od Linguistic Data Consortium (LDC), University of Pennsylvania (anotováno 60 000 slov). V současné době se připravují podklady pro analytické značkování, dále se projekt soustředí na analytické značkování a na získání podkladů pro tektogramatický popis arabské věty.
  LDC2004T23, ISBN 1-58563-319-4

- **VALLEX 1.0,** http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/
  RIV/00216208:11320/03:00002609
  Valenční slovník českých sloves, verze 1.0 je souborem lingvistických dat a dokumentace, který je výsledkem snahy o formální popis valence českých sloves. Verze 1.0 slovníku obsahuje přibližně 1400 sloves, pro něž bylo vytvořeno na 4000 valenčních rámců (1000 nejčastějších sloves z ČNK a jejich vidové protějšky). Při budování VALLEXu je kladen důraz na skutečnost, aby byl slovník snadno a rychle čitelný pro člověka, i na možnost jeho využití v automatických procedurách. Proto je slovník k dispozici v několika formátech: HTML verze (umožňuje snadnou a rychlou orientaci ve slovníku a vyhledávání podle nejrůznějších kritérií), verze pro tisk a XML verze. Po zaregistrování je pro nekomerční účely volně k využití.

- **Český anotovaný korpus,** http://ckl.mff.cuni.cz/~sgd/CAC.html.

RIV/zatím nepřiděleno
Anotovaný korpus českého jazyka (o celkovém objemu 560 000 slov) vznikl konverzí původního korpusu anotovaného v Ústavu pro jazyk český AV v sedmdesátých letech. Konverzí vnitřního kódování a anotačních schémat (na morfologické a syntakticko-analytické rovině) získáváme korpus, který je „kompatibilní" s Pražským závislostním korpusem. Byla dokončena konverze vnitřního kódování a morfologického anotování.

- **Prague Czech-English Dependency Treebank, PCEDT 1.0,**
  RIV/zatím nepřiděleno
  http://wave.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T25
  Prague Czech-English Dependency Treebank (PCEDT) je paralelní, česko-anglický závislostní korpus, který byl v roce 2004 vydán v Linguistic Data Consortium (LDC, LDC2004T25, ISBN: 1-58563-321-6). Základ paralelního korpusu tvoří překlad přibližně jedné poloviny (24 tis. vět) textů pensylvánského PennTreebanku, verze 3 (vydaného v LDC v roce 1999), který je hlavním zdrojem trénovacích a testovacích dat pro parsery angličtiny. Česká část PCEDT je automaticky morfologicky, analyticky i tektogramaticky označkována, anglická část je automaticky převedena z frázové gramatiky do závislostních analytických i tektogramatických struktur. Vzorek pětiset paralelních vět, určený pro testování, byl navíc na tektogramatické rovině anotován ručně v obou jazycích. Testovací české věty byly přeloženy čtyřmi různými překladatelskými společnostmi do angličtiny a slouží jako referenční překlady pro automatickou evaluaci výstupů překladového systému. Dále budou součástí korpusu paralelní texty z Readers' Digestu (50 tis. vět), překladový česko-anglický slovník forem, nástroje pro automatické sestavení překladového modelu z paralelních dat a nástroje pro zobrazování a vyhledávání v závislostních strukturách.

- **Czech Broadcast News Speech**, vydáno LDC, 2004
  RIV/zatím nepřiděleno
  (katalogové číslo LDC2004S01, ISBN 1- 58563-280-5)
  řečový signál: 22,05 kHz, 16 bitů
  rozsah korpusu: cca 50 hod vysílání
  stanice: ČRo1, ČRo2, ČRo3, ČTV, Prima

- **Czech Broadcast News Transcripts**, vydáno LDC, 2004
  RIV/zatím nepřiděleno
  (katalogové číslo LDC2004T01, ISBN 1-58563-281-3)

- **Korpusy spontánních promluv projektu MALACH** (ZČU Plzeň)
  - Český korpus anotovaných výpovědí lidí přeživších holocaust:
  RIV/zatím nepřiděleno
  řečový signál: 44,1 kHz
  (stereo, 1. kanál - „řečník" poskytující výpověď,
  2. kanál - moderátor), 16 bitů
  počet řečníků: 346
  rozsah korpusu: cca 100 hodin anotované řeči
  počet slov přepisu: cca 0,7 mil. slov

  - Ruský korpus anotovaných výpovědí lidí přeživších holocaust:
  RIV/zatím nepřiděleno
  řečový signál: 44,1 kHz
  (stereo, 1. kanál - „řečník" poskytující výpověď,
  2. kanál - moderátor), 16 bitů
  počet řečníků: 410
  rozsah korpusu: cca 120 hodin anotované řeči
  počet slov přepisu: cca 0,8 mil. slov

  - Slovenský korpus anotovaných výpovědí lidí přeživších holocaust (stav k 31.12.2003):
  RIV/zatím nepřiděleno

řečový signál: 44,1 kHz
(stereo, 1. kanál - „řečník“ poskytující výpověď,
2. kanál - moderátor), 16 bitů
počet řečníků: 100
rozsah korpusu: cca 25 hodin anotované řeči
počet slov přepisu: cca 0,2 mil. slov

- **Old-Church Slavonic Corpus** (OCS), http://ckl.ms.mff.cuni.cz/~ribarov.
  RIV/zatím nepřiděleno
  Korpus staroslověnských a církevněslovanských textů je vytvářen na základě dříve zpracovaných rukopisů z Ústavu pro makedonský jazyk, Skopje, Makedonie. Tento korpus obsahuje cca 600 000 slovních forem, lemmatizovaných a morfologicky označkovaných pomocí základní množiny (27) značek. Některé slovní formy (dle příslušnosti) mají asociovaný překlad, případně i referenci k jiným zdrojům. Slovní zásoba pokrývá období od 12. do cca 17. století.

Nástroje vyvíjené v rámci jednotlivých projektů Centra:

- **TrEd**
  Grafický nástroj určený k anotaci a prezentaci stromových struktur rozšiřitelný prostřednictvím uživatelem definovaných maker. Zahrnuje též nástroje pro konverze souvisejících datových formátů, dávkové zpracování souborů a na rozložení dávkového zpracování mezi skupinu výpočetních strojů. Licence GPL, http://ckl.mff.cuni.cz/~pajas/tred.

- **Nástroj pro automatický převod analytických stromových struktur na tektogramatické**
  Automatické předzpracování přechodu mezi anotací na analytické rovině k anotaci na tektogramatické rovině - soubor procedur ve formě maker pro editor TrEd. Obsahuje například algoritmy pro vypouštění uzlů funkčních slov a interpunkce, spojení analytických tvarů sloves, spojení uzlů modálních sloves s významovým slovesem, přiřazení tektogramatických lemmat uzlům, přiřazení hodnot gramatémů na základě morfologických značek z analytické roviny; http://ufal.mff.cuni.cz/publications/year2001/MN+dodat.doc.

- **XSH**
  Univerzální nástroj na interaktivní i dávkové zpracování XML souborů prostřednictvím jednoduchého jazyka založeného na standardu XPath. Licence GPL, http://xsh.sourceforge.net.

- **NetGraph**
  Souběžně s Pražským závislostním korpusem (PDT) je vyvíjen nástroj Netgraph, program pro prohledávání PDT (a jiných korpusů podobného formátu). Netgraph má architekturu klient-server a umožňuje uživatelům vyhledávat v korpusu, umístěném na výkonném serveru, z kteréhokoliv bodu internetu pomocí uživatelsky přívětivého, ale přesto velmi výkonného grafického rozhraní. Přehledný, plně grafický dotazovací jazyk je každým rokem zesilován – v roce 2003 přibyly především relace jiné než rovnítko, negace a odkazy na hodnoty atributů jiných uzlů.
  V listopadu 2003 byl Netgraph v rámci oboustranné spolupráce instalován rovněž v Linguistic Data Corporation (LDC) na University of Pennsylvania ve Philadelphii v USA, kde slouží k prohledávání arabského korpusu, tamním pracovištěm vytvářeného.
  Netgraph je pro akademické účely volně k dispozici na internetu, včetně podrobné dokumentace – viz http://quest.ms.mff.cuni.cz/netgraph.

- **Syntaktické analyzátory češtiny** ("parsery")
  V CKL se paralelně vyvíjejí nástroje pro povrchovou syntaktickou analýzu (odpovídající analytické rovině PDT) založené na různých přístupech.
  - Statistický parser (tzv. Zemanův parser)
  Tento parser je založen na statistickém modelování závislostí mezi slovy. Balíček s parserem

bude vyvěšen ke stažení na domovské stránce CKL a analýza bude také pokusně zprovozněna on-line prostřednictvím webových formulářů.

- Pravidlový parser

Tento parser je založený na automaticky získávaných pravidlech (tzv. rule-based přístup a jeho modifikace pro závislostní syntax), neobsahuje žádné před nebo post zpracování výsledných struktur.

- **Nástroje používané ve strojovém překladu**
  Nástroje jsou podrobně popsány v dokumentaci k Prague Czech-English Dependency Treebank, který bude vydán na CDROM v r. 2004 v LDC (viz výše bod 1.5).

- **Editor pro morfologickou anotaci spontánních promluv projektu MALACH**
  Vstupem editoru pro morfologickou anotaci jsou textová data zpracovaná českým morfologickým analyzátorem a taggerem. Program umožňuje snadnou vizuální kontrolu a případnou manuální korekci automaticky označkovaného textu. Jelikož byl editor vyvinut zejména pro anotaci spontánní řeči, lze v něm též opravit hovorové tvary češtiny na tvary spisovné, přičemž je současně automaticky vytvářen slovník obsahující původní nespisovné a opravené spisovné tvary.

- **Nástroj pro vytváření anotovaných korpusů ACT**
  V rámci vývoje technologií pro zpracování psaného slovanského kulturního dědictví byl za pomoci studentů vyvinut programový balík ACT (Annotated Corpora of Text) - jazykově nezávislý nástroj pro vytváření anotovaných korpusů s řadou speciálních funkcí pro zachycení jazykových víceznačností a variant. V rámci ACT je možné lemmatizovat, desambiguovat (s možností registrovat více správných variant), morfologicky značkovat, určovat reference k jiným zdrojům, určovat víceslovní celky nejrůznějších druhů, udržovat slovník lemmat, spravovat různé redakce slovníku, pracovat s překlady a asociovat text s jeho překladem. Je podporováno libovolné vyhledávání výskytů slov, včetně kontextových dotazů a předzpracovaných komplexních dotazů jako nejrůznější typy indexů, retrográdních indexů apod. V rámci ACT lze nalézt i prostředí pro zpracování lexikálních kartotéčních lístečků s cílem zpětné rekonstrukce původních excerpovaných textů. Licence GPL.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Příloha 2. Seznam publikací

2000

Začátek projektu v červenci 2000, publikace vydané v roce 2000 referovaly o práci v předcházejících projektech.

2001

1. RIV/nebylo přiděleno
   Hajič, Jan (2001): Statistické modelování a automatická analýza přirozeného jazyka (morfologie, syntax, překlad). In Slovenčina a čeština v počítačovom spracovaní (zborník referátov zo seminára Bratislava 26.-27.10.2001 (ed.A. Jarošová)) VEDA, vydavateľstvo SAV, Bratislava, ISBN 80-224-0692-9.
2. RIV/nebylo přiděleno
   Hajič, Jan; Krbec, Pavel; Oliva, Karel; Květoň, Pavel; Petkevič, Vladimír (2001): Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In Proceedings of ACL 2001 Association for Computational Linguistics.
3. RIV/00216208:11320/01:00105157
   Hajič, Jan; Vidová-Hladká, Barbora; Pajas, Petr (2001): The Prague Dependency Treebank: Annotation Structure and Support. In Proceedings of the IRCS Workshop on Linguistic Databases, pp. 105--114. University of Pennsylvania, Philadelphia, USA.
4. RIV/00216208:11320/01:00105045
   Hajičová, Eva (2001): Čeština a počítače (Abstrakt). In sborník ke konferenci ZNALOSTI 2001, 19-21.6.2001 VŠE, Praha, pp. 307.
5. RIV/00216208:11320/03:00002409
   Hajičová, Eva (2001): Information Structure and Syntactic Complexity. In Proceedings of FDSL 4 Potsdam (in press).
6. RIV/00216208:11320/02:00003034
   Hajičová, Eva (2001): Possibilities and Limits of Optimality in Topic-Focus Articulation. In Current issues in formal Slavic linguistics, pp. 385--394. Peter Lang.
7. RIV/00216208:11320/01:00105669
   Hajičová, Eva (2001): Syntaktický výzkum nad Českým národním korpusem. In Čeština - univerzália a specifika 3 (eds. Z. Hladká, P. Karlík) MU Brno, ISBN 80-210-2532-8, pp. 173-181.
8. RIV/00216208:11320/01:00105329
   Hajičová, Eva; Hajič, Jan; Vidová-Hladká, Barbora; Holub, Martin; Pajas, Petr; Řezníčková, Veronika; Sgall, Petr (2001): The Current Status of the Prague Dependency Treebank. In TSD2001 Proceedings (eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer), LNAI 2166 Springer-Verlag Berlin Heidelberg New York, ISBN 3-540-42557-8, pp. 11-20.
9. RIV/nebylo přiděleno
   Hajičová, Eva; Sgall, Petr (2001): A reusable corpus needs syntactic annotations: Prague Dependency Treebank. (v tisku) Lancaster, pp.37-48.
10. RIV/00216208:11320/01:00105451
    Hajičová, Eva; Sgall, Petr (2001): Topic-focus and salience. In Proceedings of 39th Annual Meeting of the Association for Computational linguistics, 10 thconference of the European Chapter. Proceedings, pp. 268--273. Toulouse: CNRS.
11. RIV/nebylo přiděleno
    Hajičová, Eva; Sgall, Petr; Havelka, Jiří (2001): Discourse Semantics and the Salience of Referents. In Journal of Slavic Linguistics (submitted).
12. RIV/nebylo přiděleno
    Havelka, Jiří (2001): Reference and Anaphoric Relations. Studies in Linguistics and Philosophy 72, Kluwer Academic Publishers: Dordrecht, The Netherlands. ISBN 0-7923-6070-2. Review of Reference and Anaphoric Relations. Studies in Linguistics and Philosophy 72, Kluwer Academic Publishers: Dordrecht, The Netherlands. ISBN 0-7923-6070-2. In PBML 75 UK, Praha, pp. 97-100.
13. RIV/nebylo přiděleno
    Holub, Martin; Míka, Pavel (2001): MATES -- An Experimental Linguistic Database System. In Proceedings of the IRCS Workshop on Linguistic Databases, pp. 134--140. University of Pennsylvania, Philadelphia, USA.
14. RIV/nebylo přiděleno
    Ircing, P. (2001): Language Modeling of Highly Inflectional Language (Czech). PhD Study Report. Katedra kybernetiky, Centrum komputační lingvistiky, FAV ZČU, Plzeň, 32s.
15. RIV/49777513:23520/01:00064740
    Ircing, P., Psutka, J. (2001): Two-Pass Recognition of Czech Speech Using Adaptive Vocabulary. In: Text, Speech and Dialogue. The 4th International Workshop on TSD'2001, Berlin, Heidelberg, Springer-Verlag. pp.273-277.
16. RIV/nebylo přiděleno
    Jelinek, F., Byrne, W., Khudanpur, S., Hladká, B., Ney, H., Och, F.J., Curin, J., Psutka, J.: Robust Knowledge Discovery from Parallel Speech and Text Sources. In: Proceedings of the Human Language Technology Conference HLT2001, California, San Diego.

17. RIV/nebylo přiděleno
Kučerová, Ivona (2001): Teoretická lingvistika a statistické zpracování přirozeného jazyka. In sborník řady Linguae bohemicae studentinum IV. (v tisku).
18. RIV/nebylo přiděleno
Panevová, Jarmila (2001): Některé typy chyb ve stylu odborném a žurnalistickém a možnost jejich automatického odstranění. In TERMINA 2000, Sborník příspěvků z II. konference 1996 a III. konference 2000, pp. 40--47. Galén.
19. RIV/00216208:11320/01:00105809
Panevová, Jarmila (2001): Problémy reflexivního zájmena v češtině. In Přednášky z XLIV. běhu Letní školy slovanských studií (ed. J. Nehasil) UK v Praze, FF, Praha, ISBN 80-7308-004-4, pp.81-88.
20. RIV/nebylo přiděleno
Panevová, Jarmila; Hajičová, Eva; Sgall, Petr (2001): Manuál pro tektogramatické značkování (III. verze, prosinec 2001). MFF UK.
21. RIV/nebylo přiděleno
Panevová, Jarmila; Hajičová, Eva; Sgall, Petr (2001): Tectogrammatics in corpus tagging. In Perspectivs on Semantics, Pragmatics, and Discourse; A Festschrift for Ferenc Kiefer (eds (I. Kenesei, R. M. Harnish); Pragmatics and Beyond new Series, Vol.90 John Benjamins Publishing Company Amsterdam/Philadelphia, ISBN 90 272 5109 6, pp. 294-299.
22. RIV/00216208:11320/01:00105528
Peterek, Nino (2001): Recent Methods of Prosody Analysis. In PBML 76 MFF UK, Praha.
23. RIV/49777513:23520/01:00065617
Psutka, J., Ircing, P., Radová, V.: Experiments with the Recognition of Highly Inflected Spoken Language (Czech) in the Large Vocabulary Task. In: The 5th World Multiconference on Systemics, Cybernetics SCI'2001, Orlando, U.S.A., 2001, pp. 559-564.
24. RIV/00216208:11320/01:00105723
Sgall, Petr (2001): A remark on Semantics and Pragmatics in Natural Language. In PBML 76, pp. 13--22. MFF UK.
25. RIV/nebylo přiděleno
Sgall, Petr (2001): Aspect, Eventuality Types and Nominal Reference. Garland Publishing, New York - London 1999. Review of Aspect, Eventuality Types and Nominal Reference. Garland Publishing, New York - London 1999. In Slovo a slovesnost, pp. 126--130.
26. RIV/nebylo přiděleno
Sgall, Petr (2001): Etničeskij jazyk. Opyt funkcional'noj differenciacii. Specimina philologiae Slavicae, vol.121, 1999.. Review of Etničeskij jazyk. Opyt funkcional'noj differenciacii. Specimina philologiae Slavicae, vol.121, 1999.. In Slovo a slovesnost, pp. 71--74.
27. RIV/nebylo přiděleno
Sgall, Petr (2001): Functional Generative Description, Word Order and Focus. In Theoretical Linguistics 27 pp.3-19.
28. RIV/00216208:11320/01:00105499
Sgall, Petr (2001): Ohlédnutí pražského lingvisty za dvacátým stoletím. In Slovo a slovesnost 62, pp. 241--257.
29. RIV/nebylo přiděleno
Sgall, Petr (2001): Structural and Formal Linguistics in Prague (Preface). In Towards a Relational - Perspective Approach to Syntactic Semantics ISBN 7-107-14429-4, pp. xxiii-xxxxviii.
30. RIV/00216208:11320/01:00105403
Straňáková-Lopatková, Markéta (2001): Ambiguity of Prepositional Groups: Classification, Criteria and Method for Automatic Processing.. In On Prepositions (eds. L. Šaric, D. F. Reindl), Studia Slavica Oldenburgensia 8 Bibliotheks- und Informationssystem, Oldenburg, pp.263-282.
31. RIV/nebylo přiděleno
Straňáková-Lopatková, Markéta (2001): Homonymie předložkových skupin v češtině a možnost jejího automatického zpracování. MFF UK.
32. RIV/nebylo přiděleno
Straňáková-Lopatková, Markéta (2001): Některé typy syntaktické homonymie (z hlediska možnosti automatického zpracování). In Čeština - univerzália a specifika 3, Sborník konference ve Šlapanicích u Brna, 22.-24.11.2000 (eds. Z. Hladká, P. Karlík) MU Brno, ISBN 80-210-2532-8, pp. 183-195.
33. RIV/00216208:11320/01:00105021
Straňáková-Lopatková, Markéta; Kopeček, Ivan; Pala, Karel (2001): Ambiguity Problems in Human-Computer Interaction. In Proceedings of the conference UAHCI, vol.3 (ed. C. Stephanidis) LEAmahwah, New Jersey, ISBN 0-8058-3609-8, pp.486-490.
34. RIV/00216208:11320/01:00105346
Straňáková-Lopatková, Markéta; Skoumalová, Hana; Žabokrtský, Zdeněk (2001): Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In TSD2001 Proceedings (eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer), LNAI 2166 Springer-Verlag Berlin Heidelberg New York, ISBN 3-540-42557-8, pp. 142-149.
35. RIV/nebylo přiděleno
Štěpánek, Jan (2001): CD-ROM Prague Dependency Treebank 1.0. Institute of Formal and Applied Linguistics & Linguistic Data Lab. Published by Linguistic Data Consortium, University of Pennsylvania.. In PBML 76 MFF UK.
36. RIV/68378092:_____/01:38010012
Štícha, F. (2001): Kritéria gramatičnosti (Korpus jako argument a inspirace), Slovo a slovesnost, LXII, 2001, s. 161-175.
37. Uhlířová, L.: The Case of Czech possessive adjectives and their head nouns: some distributional properties. Glottometrics, č. 2, 2001, s. 1-9.
38. RIV/nebylo přiděleno
Vidová-Hladká, Barbora; Böhmová, Alena; Ribarov, Kiril (2001): Corpus Linguistics. Investigating Language Structure and Use. Cambridge Approaches to Linguistics. Cambridge University Press: Cambridge 1998. Review of Corpus Linguistics. Investigating Language Structure and Use. Cambridge Approaches to Linguistics. Cambridge University Press: Cambridge 1998. In PBML 76 MFF UK.

39. RIV/00216208:11320/01:00105047
Zeman, Daniel (2001): How Much Will a RE-based Preprocessor Help a Statistical Parser? In Proceedings of International Workshop on Parsing Technologies Tsinghua University Press, ISBN7-302-04925-4, pp.253-256.
40. RIV/00216208:11320/01:00105203
Zeman, Daniel (2001): Parsing with Regular Expressions: A Minute to Learn, a Lifetime to Master.. In PBML 75 UK, Praha, pp.29-37.
41. RIV/00216208:11320/01:00105504
Žabokrtský, Zdeněk (2001): Automatic Functor Assignment in the Prague Dependency Treebank. MFF UK.

2002

1. RIV/nebylo přiděleno
Čmejrek, Martin; Cuřín, Jan; Havelka, Jiří (2002): Czech-English Dependency-based Machine Translation: Data Preparation for the Starting up Experiments. In Prague Bulletin of Mathematical Linguistics, pp. 103--118. MFF UK.
2. RIV/00216208:11320/02:00003001
Debowski, Lukasz; Hajič, Jan; Kuboň, Vladislav (2002): Testing the Limits -- Adding a New Language to an MT System. In Prague Bulletin of Mathematical Linguistics, pp. 95--101. MFF UK.
3. RIV/00216208:11320/02:00003046
Hajič, Jan (2002): Tectogrammatical Representation: Towards a Minimal Transfer in Machine Translation. In Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), pp. 216--226. Universita di Venezia.
4. RIV/00216208:11320/02:00003028
Hajič, Jan; Čmejrek, Martin; Eisner, Jason; Penn, Gerald; Rambow, Owen; Radev, Drago; Ding, Yuan; Koo, Terry; Parton, Kristen (2002): Natural Language Generation in the Context of Machine Translation. CLSP JHU, USA.
5. RIV/00216208:11320/02:00003014
Hajič, Jan; Oard, Douglas W.; Demner-Fushman, Dina; Ramabhadran, Bhuvana; Gustman, Samuel; Byrne, William J.; Soergel, Dagobert; Dorr, Bonnie; Resnik, Philip; Picheny, Michael (2002): Cross-Language Access to Recorded Speech in the MALACH Project. In Text, Speech and Dialogue. 5th International Conference, TSD 2002, pp. 57--64. Springer.
6. RIV/49777513:23520/02:00071579
Hajič, Jan; Psutka, Josef; Ircing, Pavel; Ramabhadran, Bhuvana; Gustman, Samuel; Byrne, William J.; Psutka, Josef V.; Radová, Vlasta (2002): Automatic Transcription of Czech Language Oral History in the MALACH Project: Resources and Initial Experiments. In Text, Speech and Dialogue. 5th International Conference, TSD 2002, pp. 253--260. Springer.
7. RIV/nebylo přiděleno
Hajičová, Eva (2002): řada hesel publikace. In Encyklopedický slovník češtiny Lidové noviny.
8. RIV/00216208:11320/02:00003061
Hajičová, Eva (2002): Recenze knihy: Studie z korpusové lingvistiky. Review of Studie z korpusové lingvistiky. In Slovo a slovesnost, pp. 65--68.
9. RIV/00216208:11320/02:00003051
Hajičová, Eva (2002): Theoretical description of language as a basis of corpus annotation: The case of Prague Dependency Treebank. In Prague Linguistic Circle Papers, pp. 111--127. John Benjamins.
10. RIV/nebylo přiděleno
Hajičová, Eva; Kučerová, Ivona (2002): Argument/Valency Structure in PropBank, LCS Database and Prague Dependency Treebank: A Comparative Pilot Study. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 846--851. ELRA.
11. RIV/00216208:11320/02:00003013
Hajičová, Eva; Pajas, Petr; Veselá, Kateřina (2002): Corpus Annotation on the Tectogrammatical Layer: Summarizing the First Stages of Evaluations. In Prague Bulletin of Mathematical Linguistics, pp. 5--18. MFF UK.
12. RIV/00216208:11320/02:00003009
Hajičová, Eva; Sgall, Petr (2002): Are Linguistic Frameworks Comparable? In Computational Linguistics for the New Millennium: Divergence or Synergy?, pp. 113--122. Peter Lang.
13. RIV/00216208:11320/02:00003017
Hajičová, Eva; Sgall, Petr (2002): Dependency syntax in Functional Generative Desription. In Festschrift for P. Hellwig
14. RIV/00216208:11320/02:00003027
Hana, Jiří; Hanová, Hana; Hajič, Jan; Vidová-Hladká, Barbora; Jeřábek, Emil (2002): Manual for Morphological Annotation. MFF UK.
15. RIV/00216208:11320/02:00003064
Holub, Martin (2002): Word Frequency Distributions. Review of Word Frequency Distributions. In: Text, Speech and Language Technology, Volume 18, 2001. ISBN 0-7923-7017-1. In Prague Bulletin of Mathematical Linguistics, pp. 113--116. MFF UK.
16. RIV/00216208:11320/02:00003007
Homola, Petr (2002): Machine translation among Slavic languages. In WDS 2002, pp. 39--43. MATFYZPRESS.
17. RIV/00216208:11320/02:00003006
Honetschläger, Václav (2002): Analytical and Tectogrammatical Syntactic Parsing. In WDS 2002, pp. 33--38. MATFYZPRESS.
18. RIV/nebylo přiděleno
Kučerová, Ivona (2002): Subjekt-predikátová shoda v češtine: univerzální, nebo specifická jazyková forma? In Čeština -- univerzália a specifika, pp. x01--x10. Lidové noviny.

19. RIV/00216208:11320/02:00003054
Kučerová, Ivona; Žabokrtský, Zdeněk (2002): Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees. In Prague Bulletin of Mathematical Linguistics, pp. 77--94. MFF UK.
20. RIV/00216208:11320/02:00003058
Lopatková, Markéta; Řezníčková, Veronika; Žabokrtský, Zdeněk (2002): Valency Lexicon for Czech: from Verbs to Nouns. In Text, Speech and Dialogue. 5th International Conference, TSD 2002, pp. 147--150. Springer.
21. RIV/nebylo přiděleno
Lopatková, Markéta; Žabokrtský, Zdeněk; Skwarska, Karolina; Benešová, Václava (2002): Tektogramaticky anotovaný valenční slovník českých sloves. MFF UK.
22. RIV/00216208:11320/02:00003029
Mírovský, Jiří; Ondruška, Roman (2002): NetGraph System: Searching through the Prague Dependency Treebank. In Prague Bulletin of Mathematical Linguistics, pp. 101--104. MFF UK.
23. RIV/00216208:11320/02:00003019
Mokrý, Karel; Smrž, Otakar (2002): External Tools Not Only for ArabTeX Documents. In Proceedings of the International Symposium on the Processing of Arabic, pp. 161--165. Department of Arabic, Faculty of Arts, University of Manouba.
24. RIV/nebylo přiděleno
Oliva, Karel; Květoň, Pavel (2002): Linguistically Motivated Bigrams in Part-of-Speech Tagging of Language Corpora. In Prague Bulletin of Mathematical Linguistics, pp. 23--36. MFF UK.
25. RIV/00216208:11320/02:00003040
Ondruška, Roman; Mírovský, Jiří; Průša, Daniel (2002): Searching through Prague Dependency Treebank-Conception and Architecture. In Proceedings of The First Workshop on Treebanks and Linguistic Theories, pp. 114--122. LML, Bulgarian Academy of Sciences and SfS, Tuebingen University.
26. RIV/nebylo přiděleno
Panevová, Jarmila (2002): řada hesel publikace. In Encyklopedický slovník češtiny Lidové noviny.
27. RIV/nebylo přiděleno
Panevová, Jarmila (2002): Corpus-based Grammar or Corpus Grammar-based? In Referát přednesený na zasedání Komise pro gramatickou stavbu slovanských jazyků
28. RIV/00216208:11320/02:00003042
Panevová, Jarmila (2002): Sloveso: centrum věty; valence: centrální pojem syntaxe. In Aktuálne otázky slovenskej syntaxe, pp. x1--x5.
29. RIV/00216208:11320/02:00003000
Panevová, Jarmila (2002): Towards a Relational - Perspective Approach to Syntactic Semantics. Review of Towards a Relational - Perspective Approach to Syntactic Semantics. Peking 2001, ISBN 7-107-14429-4, 289 pp.. In Prague Bulletin of Mathematical Linguistics, pp. 133--134. MFF UK.
30. RIV/nebylo přiděleno
Panevová, Jarmila; Hajičová, Eva; Sgall, Petr (2002): Úvod do teoretické a počítačové lingvistiky I. -- Teoretická lingvistika. Karolinum.
31. RIV/00216208:11320/02:00003023
Panevová, Jarmila; Hajičová, Eva; Sgall, Petr (2002): K nové úrovni bohemistické práce: Využití anotovaného korpusu. Část 1. In Slovo a slovesnost, pp. 161--177.
32. RIV/00216208:11320/02:00003024
Panevová, Jarmila; Hajičová, Eva; Sgall, Petr (2002): K nové úrovni bohemistické práce: Využití anotovaného korpusu. Část 2. In Slovo a slovesnost, pp. 241--262.
33. RIV/00216208:11320/02:00003060
Panevová, Jarmila; Ribarov, Kiril (2002): Za poleznosta na elektronskite jazični korpusi (vrz primerot na eden tip na imenskata fraza vo češkiot jazik). In Slavistički studii, pp. 307--316. Univerzitet Sv. Kiril i Metodij.
34. RIV/00216208:11320/02:00003048
Panevová, Jarmila; Řezníčková, Veronika; Urešová, Zdeňka (2002): The Theory of Control Applied to the Prague Dependency Treebank (PDT). In Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Frameworks (TAG+6), pp. 175--180. Universita di Venezia.
35. RIV/00216208:11320/02:00003045
Pecina, Pavel; Holub, Martin (2002): Sémanticky signifikantní kolokace. MFF UK.
36. RIV/00216208:11320/02:00003016
Plátek, Martin; Gramatovici, Radu (2002): D-trivial Dependency Grammars with Global Word-Order Restrictions. MFF UK.
37. RIV/nebylo přiděleno
Podveský, Petr (2002): Finite-state machines in speech recognition. In WDS 2002, pp. 27--32. MATFYZPRESS.
38. RIV/nebylo přiděleno
Pravdová, Markéta (2002): K povaze reklamního diskurzu. In Naše řeč, pp. 177--189.
39. RIV/nebylo přiděleno
Pravdová, Markéta (2002): McSvět a místo člověka v něm. In Studentská vědecká konference v Praze, pp. 418--431. Matfyzpress, UK Praha.
40. RIV/nebylo přiděleno
Pravdová, Markéta (2002): Reklama jako zvláštní typ sdělování. In Sborník ze 3. mezinárodního setkání mladých lingvistů
41. RIV/49777513:23520/02:00071605
Psutka, Josef; Ircing, Pavel (2002): Lattice Rescoring in Czech LVCSR System Using Linguistic Knowledge. In International Workshop Speech and Computer SPECOM 2002, pp. 23--26.
42. RIV/00216208:11320/02:00003031
Ribarov, Kiril (2002): Old Sources and Modern Procedures: Computer Processing of Old-Church Slavonic. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 1622--1626. European Language Resources Association.

43. RIV/00216208:11320/02:00003032
Ribarov, Kiril (2002): On the Rule-Based Parsing of Czech. In Prague Bulletin of Mathematical Linguistics, pp. 77--99. MFF UK.
44. RIV/00216208:11320/02:00003039
Ribarov, Kiril; Smrž, Otakar (2002): Searching for non-linearities in natural language. In 7th Experimental Chaos Conference Abstract Booklet, pp. 63--63. UCSD.
45. RIV/00216208:11320/02:00003033
Řezníčková, Veronika (2002): PDT: Two Steps in Tectogrammatical Annotation with respect to some Issues of Deletion. In Prague Bulletin of Mathematical Linguistics, pp. 37--52. MFF UK.
46. RIV/00216208:11320/02:00003063
Řezníčková, Veronika; Blažek, David (2002): Recenze: Beiträge der Europäischen Slavistischen Linguistik. Review of Beiträge der Europäischen Slavistischen Linguistik. Polyslav 4. Verlag Otto Sagner, München 2001. 292 p.. In Slovo a slovesnost, pp. 227--232.
47. RIV/00216208:11320/02:00003025
Řezníčková, Veronika; Urešová, Zdeňka (2002): K syntaktické anotaci textu z Českého národního korpusu: od analytické k tektogramatické rovině. In Aktuálne otázky slovenskej syntaxe
48. RIV/67985807:_____/03:06030119
Savický, Petr; Hlaváčová, Jaroslava (2002): Measures of Word Commonness. In Journal of Quantitative Linguistics, pp. 215--231. Swets & Zeitlinger.
49. RIV/nebylo přiděleno
Sgall, Petr (2002): Moravská a pražská (malostranská) koncepce aktuálního členění. In Čeština -- univerzália a specifika, pp. 51--58. Lidové noviny.
50. RIV/00216208:11320/02:00003043
Sgall, Petr (2002): Spoken Czech revisited. In Where One's Tongue Rules Well. A Festschrift for Charles E. Townsend, pp. 299--309. Slavica Publishers.
51. RIV/00216208:11320/02:00003049
Sgall, Petr (2002): The freedom of language. In Prague Linguistic Circle Papers, pp. 309--329. John Benjamins.
52. RIV/00216208:11320/02:00003056
Sgall, Petr (2002): Underlying Structures in Annotating Czech National Corpus. In Current issues in formal Slavic linguistics, pp. 499--505. Peter Lang (2001).
53. RIV/00216208:11320/02:00003050
Sgall, Petr; Böhmová, Alena (2002): The Simple Core and the Complex Periphery of Natural Language -- a Formal and a Computational View. In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), pp. 925--931. Morgan Kaufmann Publishers.
54. RIV/00216208:11320/02:00003004
Sgall, Petr; Žabokrtský, Zdeněk; Džeroski, Sašo (2002): A Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 1513--1520. ELRA.
55. RIV/00216208:11320/02:00003035
Smrž, Otakar; Šnaidauf, Jan; Zemánek, Petr (2002): Prague Dependency Treebank for Arabic: Multi-Level Annotation of Arabic Corpus. In Proceedings of the International Symposium on the Processing of Arabic, pp. 147--155. Department of Arabic, Faculty of Arts, University of Manouba.
56. RIV/00216208:11320/02:00003041
Smrž, Otakar; Zemánek, Petr (2002): Sherds from an Arabic Treebanking Mosaic. In Prague Bulletin of Mathematical Linguistics, pp. 63--76. MFF UK.
57. RIV/nebylo přiděleno
Straňáková-Lopatková, Markéta; Žabokrtský, Zdeněk (2002): Valenční slovník stokrát jinak: co je pod povrchem? In Čeština -- univerzália a specifika, pp. 361--363. Lidové noviny.
58. RIV/00216208:11320/02:00003057
Straňáková-Lopatková, Markéta; Žabokrtský, Zdeněk (2002): Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 949--956. ELRA.
59. RIV/00216208:11320/02:00003037
Štěpánek, Jan (2002): Building on Frege. Review of Building on Frege. In Prague Bulletin of Mathematical Linguistics, pp. 139--142. MFF UK.
60. RIV/nebylo přiděleno
Štícha, František (2002): Čas slovesný. Diateze. Gramatičnost. Hierarchizace sémantické struktury. Rod slovesný. Způsob slovesný. Osoba. In Encyklopedický slovník češtiny Lidové noviny.
61. RIV/00216208:11210/03:00008453
Štícha, František (2002): Česko-německá srovnávací gramatika. Argo.
62. RIV/nebylo přiděleno
Štícha, František (2002): Recenze: Český národní korpus. Úvod a příručka uživatele.. Review of Český národní korpus. Úvod a příručka uživatele. FF UK, 2000. In Slovo a slovesnost, pp. 73--74.
63. RIV/nebylo přiděleno
Uhlířová, Ludmila (2002): E-mail as a new electronic medium in Prague Language Consulting Services. In Referát přednesený na zasedání Komise pro gramatickou stavbu slovanských jazyků
64. RIV/nebylo přiděleno
Uhlířová, Ludmila (2002): Jazyková poradna v měnící se komunikační situaci u nás. In Sociologický časopis, pp. 443--455.

65. RIV/00216208:11320/02:00003036
Vidová-Hladká, Barbora (2002): Pražský závislostní korpus aneb Co tady před padesáti lety nebylo. In Pokroky matematiky, fyziky a astronomie, pp. 298--306. JCMF, Prague.
66. RIV/00216208:11320/02:00003038
Vidová-Hladká, Barbora; Ribarov, Kiril (2002): Exploring Textual Data. Review of Exploring Textual Data. In: Text, Speech and Language Technology series, volume 4. Kluwer Academic Publishers. 1998. In Prague Bulletin of Mathematical Linguistics, pp. 135--137. MFF UK.
67. RIV/00216208:11320/02:00003012
Zeman, Daniel (2002): Can Subcategorization Help a Statistical Dependency Parser? In Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), pp. 1156--1162. Morgan Kaufmann Publishers.
68. RIV/00216208:11320/02:00003020
Zeman, Daniel (2002): How to Decrease the Performance of a Statistical Parser. In Prague Bulletin of Mathematical Linguistics, pp. 53--62. MFF UK.

2003

1. RIV/nebylo přiděleno
Bering, Christian; Drożdżyński, Witold; Erbach, Gregor; Guasch, Clara; Homola, Petr; Lehmann, Sabine; Li, Hong; Krieger, Hans-Ulrich; Piskorski, Jakub; Schäfer, Ulrich; Shimada, Atsuko; Siegel, Melanie; Xu, Feiyu; Ziegler-Eisele, Dorothee (2003): Corpora and evaluation tools for multilingual names entity grammar development. In Proceedings of Multilingual Corpora Workshop at Corpus Linguistics, pp. !!!. (in press).
2. RIV/00216208:11320/03:00002964
Böhmová, Alena; Hajičová, Eva (2003): Large Language Data and the Degrees of Automation. In Proceedings of XVII International Congress of Linguists, CD-ROM, pp. x1-x6. Matfyzpress, MFF UK.
3. RIV/00216208:11320/03:00002552
Bojar, Ondřej (2003): AX - Systém pro automatizovanou extrakci lexikálně-syntaktických údajů. In MIS 2003, pp. 15--24. MATFYZPRESS.
4. RIV/00216208:11320/03:00002558
Bojar, Ondřej (2003): Building Subcorpora Suitable for Extraction of Lexico-Syntactic Information. In Proceedings of the Student Session, ESSLLI, pp. 25--34.
5. RIV/00216208:11320/03:00002534
Bojar, Ondřej; Brom, Cyril; Hladík, Milan; Vejlupek, Mikuláš; Toman, Vojtěch; Voňka, David (2003): ENTI -- Simulátor přirozeného prostředí lidského světa. In MIS 2003, pp. 3--14. MATFYZPRESS.
6. RIV/00216208:11320/03:00002517
Camuglia, Monia; Ribarov, Kiril (2003): Old-Church Slavonic in Codes. In Computational Approaches to the study of Early and Modern Slavic Languages and Texts -- Proceeedings of the Electronic Description and Edition of Slavic Sources, pp. 201--204.
7. RIV/00216208:11320/02:00003015
Čmejrek, Martin; Cuřín, Jan; Havelka, Jiří (2003): Czech-English Dependency-based Machine Translation. In EACL 2003 Proceedings of the Conference, pp. 83--90. Association for Computational Linguistics.
8. RIV/00216208:11320/03:00002955
Čmejrek, Martin; Cuřín, Jan; Havelka, Jiří (2003): Treebanks in Machine Translation. In Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 209--212. Vaxjo University Press.
9. RIV/00216208:11320/03:00002235
Drożdżyński, Witold; Homola, Petr; Piskorski, Jakub; Zinkevičius, Vytautas (2003): Adapting SProUT to processing Baltic and Slavonic languages. In Proceedings of Information Extraction for Slavonic and other Central and Eastern European Languages, pp. !!!.
10. RIV/00216208:11320/03:00002976
Gramatovici, Radu (2003): On the Recognition Power of Non-Expansive Go-Through Automata. In Annals of Bucharest University, pp. 45--54.
11. RIV/00216208:11320/03:00002565
Hajič, Jan; Homola, Petr; Kuboň, Vladislav (2003): A Simple Multilingual Machine Translation System. In Proceedings of Machine Translation Summit IX, pp. 157--164.
12. RIV/00216208:11320/03:00002696
Hajič, Jan; Honetschläger, Václav (2003): Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. In Prague Bulletin of Mathematical Linguistics, pp. 61--86. MFF UK.
13. RIV/00216208:11320/03:00002564
Hajič, Jan; Kuboň, Vladislav (2003): Tagging as a Key to Successful MT. In MIS 2003, pp. 56--65. MATFYZPRESS.
14. RIV/00216208:11320/03:00002419
Hajič, Jan; Panevová, Jarmila; Urešová, Zdeňka; Bémová, Alevtina; Kolářová, Veronika; Pajas, Petr (2003): PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 57--68. Vaxjo University Press.
15. RIV/nebylo přiděleno
Hajič, Jan; Psutka, Josef; Ircing, Pavel; Byrne, William; Mírovský, Jiří; Ramabhadran, Bhuvana; Gustman, Samuel; Psutka, Josef V.; Radová, Vlasta (2003): Language Model Data Selection for Czech ASR in the MALACH Project. In ICASSP 2003, pp. !!!. (submitted).
16. RIV/00216208:11320/03:00002445
Hajič, Jan; Urešová, Zdeňka (2003): Linguistic Annotation: from Links to Cross-Layer Lexicons. In Proceedings of The Second Workshop on Treebanks and Linguistic Theories, pp. 69--80. Vaxjo University Press.

17. RIV/00216208:11320/03:00002965
Hajičová, Eva (2003): Aspects of discourse structure. In Natural language processing between linguistic inquiry and system engineering, pp. 47--54. Editura Universitatii Alexandru Ioan Cuza.
18. RIV/00216208:11320/03:00002966
Hajičová, Eva (2003): Contextual boundness and discourse patterns. In Proceedings of XVII International Congress of Linguists, CD-ROM, pp. x1-x7. Matfyzpress, MFF UK.
19. RIV/00216208:11320/03:00002409
Hajičová, Eva (2003): Information structure and syntactic complexity. In Investigations into formal Slavic linguistics, pp. 169--180. Peter Lang.
20. RIV/00216208:11320/03:00002962
Hajičová, Eva (2003): Syntactic theory and corpus annotation need each other. In Zbornik povzetkov, 13. mednarodni slavistični kongres, 2. del, pp. 289. Medninarodni slavistični komite.
21. RIV/00216208:11320/03:00002971
Hajičová, Eva (2003): Topic-focus articulation in the Czech National Corpus. In Language and function. To the memory of Jan Firbas, pp. 185--194. John Benjamins.
22. RIV/nebylo přiděleno
Hajičová, Eva; Havelka, Jiří; Sgall, Petr (2003): Discourse Semantics and the Salience of Referents. In Journal of Slavic Linguistics, pp. 127-140.
23. RIV/nebylo přiděleno
Hajičová, Eva; Sgall, Petr (2003): Dependency syntax in Functional Generative Description. In Dependenz und Valenz -- Dependency and Valency, pp. 570--592. Walter de Gruyter.
24. RIV/00216208:11320/03:00002196
Hajičová, Eva; Sgall, Petr (2003): Information Structure, Translation and Discourse. In Textologie und Translation, pp. 107--123. Gunter Narr.
25. RIV/00216208:11320/02:00003052
Hajičová, Eva; Sgall, Petr; Buráňová, Eva (2003): Topic-Focus Articulation and degrees of salience in the Prague Dependency Treebank. In Formal Approaches to Function in Grammar. In honor of Eloise Jelinek, Arizona, pp. 165--177. John Benjamins.
26. RIV/00216208:11320/02:00003021
Hajičová, Eva; Sgall, Petr; Veselá, Kateřina (2003): Information structure and contrastive topic. In Formal approaches to Slavic linguistics. The Amherst Meeting 2002, pp. 219--234. Michigan Slavic Publications.
27. RIV/00216208:11320/03:00002572
Holan, Tomáš; Kuboň, Vladislav; Plátek, Martin; Oliva, Karel (2003): A Theoretical Basis of an Architecture of a Shell of a Reasonably Robust Syntactic Analyser. In Proceedings of Text, Speech and Dialogue 2003, pp. 58--65. Springer.
28. RIV/00216208:11320/03:00002936
Holub, Martin (2003): A New Approach to Conceptual Document Indexing: Building a Hierarchical System of Concepts Based on Document Clusters. In ISICT 2003 Proceedings of the International Symposium on Information and Communication Technologies, pp. 311--316. Trinity College Dublin.
29. RIV/00216208:11320/03:00002899
Holub, Martin; Straňák, Pavel (2003): Approaches to Building Semantic Lexicons. In WDS'03 Proceedings of Contributed Papers, Part I, pp. 173--178. MATFYZPRESS.
30. RIV/nebylo přiděleno
Homola, Petr; Rimkutė, Erika (2003): Shallow machine translation - in between of two extremes. In Proceedings of The Fifth International Tbilisi Symposium on Language, Logic and Computation, pp. !!!. (in press).
31. RIV/00216208:11320/03:00002957
Honetschläger, Václav (2003): Using a Czech Valency Lexicon for Annotation Support. In Proceedings of Text, Speech and Dialogue 2003, pp. 120--126. Springer.
32. RIV/49777513:23520/03:00000158
Ircing, Pavel; Psutka, Josef (2003): Fitting Class-Based Language Models into Weighted Finite-State Transducer Framework. In EUROSPEECH 2003 Proceedings (8th European Conference on Speech Communication and Technology), pp. 1873--1876. ISCA.
33. RIV/00216208:11320/03:00002871
Kocanda, Jiří (2003): Statistical Parsing. In WDS'03 Proceedings of Contributed Papers, Part I, pp. 161--166. MATFYZPRESS.
34. RIV/00216208:11320/03:00002967
Krbec, Pavel; Podveský, Petr; Hajič, Jan (2003): Combination of a Hidden Tag Model and a Traditional N-gram Model: A Case Study in Czech Speech Recognition. In EUROSPEECH 2003 Proceedings (8th European Conference on Speech Communication and Technology), pp. 2289--2291. ISCA.
35. RIV/00216208:11320/03:00002646
Kuboň, Vladislav (2003): Multilingual Aspects of Monolingual Corpora. In In the proceedings of Sprachtechnlogie fuer die Multilinguale Kommunikation, GLDV-Fruejahrstagung 2003, pp. 283--298. Gardez-Verlag.
36. RIV/nebylo přiděleno
Kučerová, Ivona; Řezníčková, Veronika (2003): Korpus jako výzva k syntaktické analýze. Poznámky k syntaktické derivaci deverbativních substantiv v češtině. In Slavia, pp. 267--274.
37. RIV/00216208:11320/03:00002961
Kučová, Lucie; Kolářová, Veronika; Žabokrtský, Zdeněk; Pajas, Petr; Čulo, Oliver (2003): Anotování koreference v Pražském závislostním korpusu. MFF UK.
38. RIV/nebylo přiděleno
Kupera, Břetislav (2003): Genetic Algorithms and Artificial Neural Network in Natural Language Processing. In WDS'03 Proceedings of Contributed Papers, Part I, pp. 156--160. MATFYZPRESS.

39. RIV/00216208:11320/03:00002398
Květoň, Pavel (2003): Language for Grammatical Rules. MFF UK.
40. RIV/nebylo přiděleno
Lopatková, Markéta (2003): Issue of Valency in Prague Dependency Treebank: Creating valency lexicon of Verbs. (Abstract) . In XVII International Congress of Linguists Abstracts , pp. 153-153. MFF UK.
41. RIV/00216208:11320/03:00002313
Lopatková, Markéta (2003): O homonymii předložkových skupin v češtine (Co umí počítač?) Karolinum.
42. RIV/00216208:11320/03:00002804
Lopatková, Markéta (2003): Valency in the Prague Dependency Treebank: Building the Valency Lexicon. In Prague Bulletin of Mathematical Linguistics, pp. 37--60. MFF UK.
43. RIV/00216208:11320/03:00002609
Lopatková, Markéta; Žabokrtský, Zdeněk; Skwarska, Karolina; Benešová, Václava (2003): VALLEX 1.0 Valency Lexicon of Czech Verbs. MFF UK.
44. RIV/00216208:11320/03:00002969
Oliva, Karel; Květoň, Pavel; Ondruška, Roman (2003): The Computational Complexity of Rule-Based Part-of-Speech Tagging. In Proceedings of Text, Speech and Dialogue 2003, pp. 82--89. Springer.
45. RIV/00216208:11320/03:00002968
Ondruška, Roman; Panevová, Jarmila; Štěpánek, Jan (2003): An Exploitation of the Prague Dependency Treebank: A Valency Case. In Proceedings of the Workshop on Shallow Processing of Large Corpora (SproLaC 2003), pp. 69--77. UCREL, Lancaster University.
46. RIV/nebylo přiděleno
Panevová, Jarmila (2003): Existuje chyba v syntaxi? In Sborník prací Filozoficko-přírodovědecké fakulty Slezské univerzity v Opavě, pp. 145--153. Slezská univerzita v Opavě.
47. RIV/nebylo přiděleno
Panevová, Jarmila (2003): Some Issues of Syntax and Semantics of Verbal Modifications. In Proceedings MTT 2003, First International Conference on Meaning-Text Theory, pp. 139--146. Ecole Normale Supérieure.
48. RIV/00216208:11320/03:00002958
Plátek, Martin; Lopatková, Markéta; Oliva, Karel (2003): Restarting Automata: Motivations and Applications. In Proceedings of the workshop Petrinetze, pp. 90--96. Technische Universitaet Muenchen.
49. RIV/49777513:23520/03:00000156
Psutka, Josef; Iljuchin, Ilja; Ircing, Pavel; Psutka, Josef V.; Trejbal, Václav; Byrne, William J.; Hajič, Jan; Gustman, Samuel (2003): Building LVCSR System for Transcription of Spontaneously Pronounced Russian Testimonies in the MALACH Project: Initial Steps and First Results. In Proceedings of Text, Speech and Dialogue 2003, pp. 327--332. Springer.
50. RIV/49777513:23520/03:00000157
Psutka, Josef; Ircing, Pavel; Psutka, Josef V.; Radová, Vlasta; Byrne, William; Hajič, Jan; Mírovský, Jiří; Gustman, Samuel (2003): Large Vocabulary ASR for Spontaneous Czech in the MALACH Project. In EUROSPEECH 2003 Proceedings (8th European Conference on Speech Communication and Technology), pp. 1821--1824. ISCA.
51. RIV/49777513:23520/03:00000155
Psutka, Josef; Ircing, Pavel; Psutka, Josef V.; Radová, Vlasta; Byrne, William J.; Venkataramani, Veera; Hajič, Jan; Gustman, Samuel (2003): Towards Automatic Transcription of Spontaneous Czech Speech in the MALACH Project. In Proceedings of Text, Speech and Dialogue 2003, pp. 214--219. Springer.
52. RIV/00216208:11320/03:00002681
Rambow, Owen; Dorr, Bonnie; Kipper, Karin; Kučerová, Ivona; Palmer, Martha (2003): Automatically Deriving Tectogrammatical Labels from Other Resources: A Comparison of Semantic Labels Across Frameworks. In Prague Bulletin of Mathematical Linguistics, pp. 23--35. MFF UK.
53. RIV/00216208:11320/03:00002526
Ribarov, Kiril; Camuglia, Monia (2003): Incorporation of Old-Church Slavonic Card-Files into a Corpus. In Scripta & e-Scripta, pp. 65--74. Institute of Literature, Bulgarian Academy of Sciences.
54. RIV/nebylo přiděleno
Řezníčková, Veronika (2003): Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora. In Proceedings of the Workshop on Shallow Processing of Large Corpora (SProLaC 2003), pp. 88--97. UCREL, Lancaster University.
55. RIV/00216208:11320/03:00002893
Semecký, Jiří (2003): Semantic Word Classes Extracted from Text Clusters. In WDS'03 Proceedings of Contributed Papers, Part I, pp. 167--172. MATFYZPRESS.
56. RIV/00216208:11320/02:00003018
Sgall, Petr (2003): Dynamics in the meaning of the sentence and of discourse. In Meaning: The Dynamic Turn, pp. 169--184. Elsevier Science Ltd..
57. RIV/00216208:11320/03:00002163
Sgall, Petr (2003): From Data to Speech. Language Generation in Context. Review of From Data to Speech. Language Generation in Context. In Journal of Pragmatics, pp. 315--319. Elsevier.
58. RIV/00216208:11320/03:00002972
Sgall, Petr (2003): From functional sentence perspective to topic-focus articulation. In Language and function. To the memory of Jan Firbas, pp. 279--287. John Benjamins.
59. RIV/nebylo přiděleno
Sgall, Petr (2003): Introductory remarks (to the Workshop on Discourse Patterns). In Proceedings of XVII International Congress of Linguists, CD-ROM, pp. x1-x5. Matfyzpress, MFF UK.
60. RIV/00216208:11320/03:00002234
Sgall, Petr (2003): Lingvistické ohlédnutí za dvacátým stoletím. In Český jazyk a literatura, pp. 157--164. SPN & Fortuna.

61. RIV/00216208:11320/03:00002413
    Sgall, Petr (2003): Slavistics and the history of topic-focus studies. In Investigations into formal Slavic linguistics, pp. 201--212. Peter Lang.
62. RIV/00216208:11320/02:00003053
    Sgall, Petr (2003): Topic-Focus Articulation in Corpus Annotation. In Natural language processing between linguistic inquiry and system engineering, pp. 95--101. Editura Universitatii Alexandru Ioan Cuza.
63. RIV/00216208:11210/03:00008453
    Štícha, František (2003): Česko-německá srovnávací gramatika. Argo.
64. RIV/00216208:11210/03:00008452
    Štícha, František (2003): Gramatický výzkum dříve a dnes: korpus jako výzva. In Tradícia a perspektívy gramatického výskumu na slovensku. Veda, pp. 24-31.
65. RIV/nebylo přiděleno
    Uhlířová, Ludmila (2003): The Czech determiners tento 'this' and ten 'that' in discourse structure. In CD CIL17
66. RIV/nebylo přiděleno
    Uhlířová, Ludmila (2003): This/that in discourse structure: An evidence from the Czech National Corpus. In 36th International Meeting of the Societas Linguistica Europaea
67. RIV/nebylo přiděleno
    Uhlířová, Ludmila (2003): Zipf´s law for pairs of words. In Journal of Quantitative Linguistics (in press)
68. RIV/00216208:11320/03:00002347
    Veselá, Kateřina; Havelka, Jiří (2003): Anotování aktuálního členění věty v Pražském závislostním korpusu. MFF UK.
69. RIV/00216208:11320/03:00002963
    Veselá, Kateřina; Peterek, Nino; Hajičová, Eva (2003): Some observations on contrastive topic in Czech spontaneous speech. In Proceedings of XVII International Congress of Linguists, CD-ROM, pp. !!!. Matfyzpress, MFF UK.
70. RIV/00216208:11320/03:00002686
    Veselá, Kateřina; Peterek, Nino; Hajičová, Eva (2003): Topic-Focus Articulation in PDT: Prosodic Characteristics of Contrastive Topic. In Prague Bulletin of Mathematical Linguistics, pp. 5--22. MFF UK.
71. RIV/00216208:11320/03:00002102
    Žabokrtský, Zdeněk (2003): Word Sense Disambiguation. The Case for Combinations of Knowledge Sources. Review of Word Sense Disambiguation. The Case for Combinations of Knowledge Sources. CSLI Publications, 2003. Stanford California. ISBN 1-57586-390-1 (pbk.), 1-57586-389-8 (hard). Pp. xvi+175. In Prague Bulletin of Mathematical Linguistics, pp. 151--153. MFF UK.
72. RIV/00216208:11320/03:00002975
    Žabokrtský, Zdeněk; Smrž, Otakar (2003): Arabic Syntactic Trees: from Constituency to Dependency. In EACL 2003 Conference Companion, pp. 183--186. Association for Computational Linguistics.

2004

1. RIV/zatím nebylo přiděleno
   Bojar, Ondřej (2004): Problems of Inducing Large Coverage Constraint-Based Dependency Grammar for Czech. In Proceedings of International Workshop on Constraint Solving and Language Processing, CSLP 2004, pp. 29--42. Roskilde University.
2. RIV/zatím nebylo přiděleno
   Bojar, Ondřej; Benešová, Václava (2005): VALEVAL: Recent Experiments with the Valency Lexicon of Czech Verbs. In submitted to Verb Workshop 2005 (submitted)
3. RIV/zatím nebylo přiděleno
   Byrne, William J.; Doermann, David; Franz, Martin; Gustman, Samuel; Hajič, Jan; Oard, Douglas W.; Picheny, Michael; Psutka, Josef V.; Ramabhadran, Bhuvana; Soergel, Dagobert; Ward, Todd; Zhu, Wang (2004): Automatic Recognition of Spontaneous Speech for Access to Multilingual Oral History Archives. In IEEE Transactions on Speech and Audio Processing, pp. 420-435.
4. RIV/zatím nebylo přiděleno
   Cinková, Silvie (2004): Extraction of Swedish Verb-Noun Collocations from a Large Msd-Annotated Corpus. In The Prague Bulletin of Mathematical Linguistics 82, pp. 99--102.
5. RIV/zatím nebylo přiděleno
   Cinková, Silvie (2004): Manuál pro tektogramatickou anotaci angličtiny. In ÚFAL/CKL, pp. 2-172.
6. RIV/zatím nebylo přiděleno
   Cinková, Silvie (2004): Recenze - Ruslan Mitkov (ed.) The Oxford Handbook of Computational Linguistics. In The Prague Bulletin of Mathematical Linguistics 82, pp. 87--94.
7. RIV/zatím nebylo přiděleno
   Cinková, Silvie; Kolářová, Veronika (2004): Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In Korpusy a korpusová lingvistika v zahraničí a na Slovensku (in press)
8. RIV/zatím nebylo přiděleno
   Cuřín, Jan; Čmejrek, Martin; Havelka, Jiří; Hajič, Jan; Kuboň, Vladislav; Žabokrtský, Zdeněk (2004): Prague Czech-English Dependency Treebank Version 1.0. In Linguistic Data Consortium (LDC) Linguistic Data Consortium (LDC).
9. RIV/zatím nebylo přiděleno
   Čmejrek, Martin; Cuřín, Jan; Havelka, Jiří (2004): Prague Czech-English Dependency Treebank: Any Hopes for a Common Annotation Scheme? In HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, pp. 47--54. Association for Computational Linguistics.

10. RIV/zatím nebylo přiděleno
Čmejrek, Martin; Cuřín, Jan; Havelka, Jiří; Hajič, Jan; Kuboň, Vladislav (2004): Prague Czech-English Dependency Treebank. Syntactically Annotated Resources for Machine Translation. In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1597-1600. European Language Resources Association.
11. RIV/zatím nebylo přiděleno
Frank, Anett; Semecký, Jiří (2004): . In Proceedings of the 5th International Conference on Linguistically Interpreted Corpora, LINC 2004, Proceedings of the 5th International Conference on Linguistically Interpreted Corpora, LINC 2004
12. RIV/zatím nebylo přiděleno
Guthrie, Louise; Basili, Roberto; Zanzotto, Fabio; Boncheva, Kalina; Cunningham, Hamish; Guthrie, David; Cui, Jia; Cammisa, Marco; Cheng-Chieh Liu, Jerry; Farria Martin, Cassia; Haralambiev, Kristiyan; Holub, Martin; Machery, Klaus; Jelinek, Frederick (2004): Large Scale Experiments for Semantic Labeling of Noun Phrases in Raw Text. In Proceedings of LREC 2004
13. RIV/zatím nebylo přiděleno
Hajič, Jan (2004): Complex Corpus Annotation: The Prague Dependancy Treebank. In in prep. Jazykovedný ústav Ľ. Štúra, SAV.
14. RIV/zatím nebylo přiděleno
Hajič, Jan (2004): Disambiguation of Rich Inflection (Computational Morphology of Czech). Nakladatelství Karolinum.
15. RIV/zatím nebylo přiděleno
Hajič, Jan (2004): History of Computational Linguistics. In A Companion to Digital Humanities Blackwell Publishing.
16. RIV/zatím nebylo přiděleno
Hajič, Jan; Holub, Martin; Hučínová, Marie; Pavlík, Martin; Pecina, Pavel; Straňák, Pavel; Šidák, Pavel (2004): Validating and Improving the Czech WordNet via Lexico-Semantic Annotation of the Prague Dependency Treebank. In Proceedings of LREC 2004
17. RIV/zatím nebylo přiděleno
Hajič, Jan; Panevová, Jarmila; Buráňová, Eva; Urešová, Zdeňka; Bémová, Alla; Štěpánek, Jan; Pajas, Petr; Kárník, Jiří (2004): Anotace na analytické rovině. Návod pro anotátory. In UFAL/CKL technical report MFF UK, TR-2004-23.
18. RIV/zatím nebylo přiděleno
Hajič, Jan; Smrž, Otakar; Zemánek, Petr; Pajas, Petr; Šnaidauf, Jan; Beška, Emanuel; Kráčmar, ?; Hassanová, Kamila (2004): Prague Arabic Dependency Treebank 1.0. Linguistic Data Consortium.
19. RIV/zatím nebylo přiděleno
Hajič, Jan; Smrž, Otakar; Zemánek, Petr; Šnaidauf, Jan; Beška, Emanuel (2004): Prague Arabic Dependency Treebank: Development in Data and Tools. In Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools, pp. 110--117. ELDA.
20. RIV/zatím nebylo přiděleno
Jan Hajič, Jan; Urešová, Zdena; Bémová, Alla; Kaplanová, Marie (2004) Pražský závislostní korpus. Anotace na tektogramatické rovině (úroveň 3). In UFAL/CKL technical report MFF UK, TR-2004-24.
21. RIV/zatím nebylo přiděleno
Jan Hajič, Jan; Urešová, Zdena; Bémová, Alla; Kaplanová, Marie (2004): The Prague Dependency Treebank. Annotace on tectogrammatical level. (Translated by Böhmová, Alena; Cinková, Silvie). In UFAL/CKL technical report MFF UK, TR-2004-25.
22. RIV/zatím nebylo přiděleno
Hajičová, Eva; Havelka, Jiří; Sgall, Petr (2004): Topic and focus, anaphoric relations and degrees of salience. In Prague Linguistic Circle Papers / Travaux du cercle linguistique de Prague N.S. (in press) John Benjamins.
23. RIV/zatím nebylo přiděleno
Hajičová, Eva; Havelka, Jiří; Sgall, Petr; Veselá, Kateřina; Zeman, Daniel (2004): Issues of Projectivity in the Prague Dependency Treebank. In Prague Bulletin of Mathematical Linguistics MFF UK (in press).
24. RIV/zatím nebylo přiděleno
Hajičová, Eva; Sgall, Petr (2004): Degrees of Contrast and the Topic-Focus Articulation. In Language, Context & Cognition - Information Structure - Theoretical and Empirical Aspects, pp. 1--13. Walter de Gruyter.
25. RIV/zatím nebylo přiděleno
Hajičová, Eva; Sgall, Petr (2004): Translation and Information Structure. In Neue Perspektiven in der Übersetzung- und Dolmetscherwissenschaft, pp. 235-247. AKS-Verlag.
26. RIV/zatím nebylo přiděleno
Havelka, Jiří; Hajič, Jan; Kuboň, Vladislav (2004): Prague Czech-English Dependency Treebank. Syntactically Annotated Resources for Machine Translation. In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1597-1600. European Language Resources Association.
27. RIV/zatím nebylo přiděleno
Hlaváčová, Jaroslava (2004): Automatické rozpoznávání českých derivačních předpon. In accepted for publication in proceedings CICLING 2005
28. RIV/zatím nebylo přiděleno
Hlaváčová, Jaroslava; Klímová, Jana (2004): Derivational Relations in Flectional Languages - Czech Case. In Proceeding LREC 2004, pp. 1239-1242.
29. RIV/zatím nebylo přiděleno
Holub, Martin; Diviš, Jiří; Pávek, Jan; Pecina, Pavel; Semecký, Jiří (2004): Topics of Texts. Annotation, Automatic Searching and Indexing. In UFAL/CKL technical report MFF UK, TR-2004-21.
30. RIV/zatím nebylo přiděleno
Holub, Martin; Semecký, Jiří; Diviš, Jiří (2004): Searching for Topics in a Large Collection of Texts. In Proceedings of ACL 2004

31. RIV/zatím nebylo přiděleno
Homola, Petr (2004): On some aspects on machine translation among related languages. In Proceedings of the Ninth ESSLLI Student Session
32. RIV/zatím nebylo přiděleno
Homola, Petr; Kuboň, Vladislav (2004): A translation model for languages of acceding countries. In Proceedings of the EAMT Workshop
33. RIV/zatím nebylo přiděleno
Homola, Petr; Piskorski, Jakub (2004): How can shallow NLP help a machine translation system. In Proceedings of the Conference Human Language Technologies - The Baltic Perspective
34. RIV/zatím nebylo přiděleno
Homola, Petr; Rimkutė, Erika (2004): Mašininis vertimas tarp artimų kalbų. In in press Kaunas Technology University.
35. RIV/zatím nebylo přiděleno
Homola, Petr; Tolvaj, Béla (2004): Distributed translation memories and shallow MT. In MIS 2004 MATFYZPRESS.
36. RIV/zatím nebylo přiděleno
Klusáček, David (2004): Optimal Detection in Case of the Sparse Training Data. In Proceedings of ODYSSEY04, pp. 97--104.
37. RIV/zatím nebylo přiděleno
Kolář, J.; Švec, ?; Psutka, Josef V. (2004): Automatic Punctuation Annotation in Czecch Broadcast News Speech. In Proceeding of 9th International Conference Speech and Computer, SPECOM'2004, pp. 319-325.
38. RIV/zatím nebylo přiděleno
Kuboň, Vladislav; Cuřín, Jan; Čmejrek, Martin; Havelka, Jiří (2004): Building parallel bilingual syntactically annontated corpus. In Proceedings of The First International Joint Conference on Natural Language Processing, pp. 141-146.
39. RIV/zatím nebylo přiděleno
Kučová, Lucie; Hajičová, Eva (2004): Coreferential Relations in the Prague Dependency Treebank. In Proceedings of DAARC2004, pp. 97-102.
40. RIV/zatím nebylo přiděleno
Kučová, Lucie; Hajičová, Eva (2004): Coreferential Relations in the Prague Dependency Treebank. In Sborník prací ke konferenci FDSL-5 (in press)
41. RIV/zatím nebylo přiděleno
Kučová, Lucie; Hajičová, Eva (2004): Prague Dependency Treebank: Enrichment of the Underlying Syntactic Annotation by Coreferential Mark-Up. In Prague Bulletin of Mathematical Linguistics, pp. 23-34.
42. RIV/zatím nebylo přiděleno
Lopatková, Markéta; Panevová, Jarmila (2005): Recent developments of the theory of valency in the light of the Prague Dependency Treebank.. In Sborník SNK (in press)
43. RIV/zatím nebylo přiděleno
Lopatková, Markéta; Panevová, Jarmila (2004): Valence vybraných skupin sloves (k některým slovesům dandi a recipiendi). In Čeština - univerzália a specifika, Sborník konference ve Šlapanicích U BrnaČeština - univerzália a specifika, pp. 348--356. Nakladatelství Lidové noviny.
44. RIV/zatím nebylo přiděleno
Lopatková, Markéta; Plátek, Martin; Kuboň, Vladislav (2005): Závislostní redukční analýza přirozených jazyků. In Proceedings of ITAT 2004 (in press) University of P. J. Šafařík.
45. RIV/zatím nebylo přiděleno
Lopatková, Markéta; Žabokrtský, Zdeněk (2004): Testování konzistence a úplnosti valenčního slovníku českých sloves. In Proceedings of ITAT 2003, pp. 73-82. University of P. J. Šafařík.
46. RIV/zatím nebylo přiděleno
Panevová, Jarmila (2004): Všeobecné aktanty očima Pražského závislostního korpusu (PZK). In Korpus jako zdroj dat o češtině. Sborník konference ve Šlapanicích (in press)
47. RIV/zatím nebylo přiděleno
Piskorski, Jakub; Homola, Petr; Marciniak, Małgorzata; Mykowiecka, Agnieszka; Przepiórkowski, Adam; Woliński, Marcin (2004): Information extraction for Polish using the SProUT platform. In Proceedings of the International IIS:IIPWM WM'04 Conference, pp. 227--236. Springer Verlag.
48. RIV/zatím nebylo přiděleno
Pravdová, Markéta (2004): K způsobům persvaze v reklamních projevech. In Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV), pp. 131-136. Otto Sagner.
49. RIV/zatím nebylo přiděleno
Pravdová, Markéta (2004): Reklama jako zvláštní typ sdělování. In Vztah langue a parole v perspektivě "interaktivního obratu" v lingvistickém zkoumání, pp. 248-253. UP Olomouc.
50. RIV/zatím nebylo přiděleno
Psutka, Josef V.; Hajič, Jan; Byrne, William J. (2004): The Development of ASR for Slavic Languages in the MALACH Project. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2004, pp. 749-752.
51. RIV/zatím nebylo přiděleno
Psutka, Josef V.; Ircing, Pavel; Hajič, Jan; Radová, Vlasta; Psutka, Josef V.; Byrne, William J. (2004): Issues in annotation of the Czech spontaneous speech corpus in the MALACH project. In Proceedings of the 4th International Conference on Language Resources and Evaluation LREC , pp. 607-610.
52. RIV/zatím nebylo přiděleno
Radová, Vlasta; Psutka, Josef V.; Müller, ?; Byrne, William J.; Psutka, Josef V.; Ircing, Pavel; Matoušek, ? (2004): Czech Broadcast News Speech. Linguistic Data Consortium, University of Pennsylvania.

53. RIV/zatím nebylo přiděleno
Radová, Vlasta; Psutka, Josef V.; Müller, ?; Byrne, William J.; Psutka, Josef V.; Ircing, Pavel; Matoušek, ? (2004): Czech Broadcast News Transcripts. Linguistic Data Consortium, University of Pennsylvania.
54. RIV/zatím nebylo přiděleno
Ribarov, Kiril (2004): Automatic Building of a Dependency Tree - The Rule-Based Approach and Beyond. MFF UK.
55. RIV/zatím nebylo přiděleno
Ribarov, Kiril (2004): Towards Intelligent Written Cultural Heritage Processing - Lexical Processing. In Proceeedings of LREC 2004
56. RIV/zatím nebylo přiděleno
Ribarov, Kiril; Bubník, Jiří; Čelák, Jiří; Janota, Vojtěch; Kara, Alexandr; Novák, Václav; Vondra, Tomáš (2004): ACT - Computer Processing of Written Cultural Heritage Sources. In Proceedings of INFORUM 2004 Conference
57. RIV/zatím nebylo přiděleno
Ribarov, Kiril; Bubník, Jiří; Čelák, Jiří; Janota, Vojtěch; Kara, Alexandr; Novák, Václav; Vondra, Tomáš (2004): We present the ACT Tool. In Scripta & e-Scripta Bulgarian Academy of Sciences.
58. RIV/zatím nebylo přiděleno
Sgall, Petr (2004): Co pomůže češtině. O potřebě přejít od školské spisovnosti ke standardnímu vyjadřování. In Přítomnost, pp. 52--53.
59. RIV/zatím nebylo přiděleno
Sgall, Petr (2004): K obohacování spisovné češtiny. In Čeština - univerzália a specifika, pp. 77--85. Nakladatelství Lidové noviny.
60. RIV/zatím nebylo přiděleno
Sgall, Petr (2004): Types of Languages and the Simple Pattern of the Core of Language. In Linguistics Today - Facing a Greater Challenge (Plenary lectures from CIL 17), pp. 243--265. Benjamins.
61. RIV/zatím nebylo přiděleno
Sgall, Petr; Panevová, Jarmila (2004): Jak psát a nepsat česky. In Učební texty UK v Praze Karolinum.
62. RIV/zatím nebylo přiděleno
Sgall, Petr; Panevová, Jarmila; Hajičová, Eva (2004): Deep Syntactic Annotation: Tectogrammatical Representation and Beyond. In HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, pp. 32--38. Association for Computational Linguistics.
63. RIV/zatím nebylo přiděleno
Smrž, Otakar (2004): Finite State Morphology. Review of Finite State Morphology. CSLI Publications, Stanford, California, 2003 (CSLI Studies in Computational Linguistics, xviii+510 pp and CD-ROM, ISBN 1-57586-434-7). In Prague Bulletin of Mathematical Linguistics MFF UK (in press).
64. RIV/zatím nebylo přiděleno
Smrž, Otakar; Pajas, Petr (2004): MorphoTrees of Arabic and Their Annotation in the TrEd Environment. In Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools, pp. 38--41. ELDA.
65. RIV/zatím nebylo přiděleno
Štícha, František (2004): Nominativ a instrumentál predikátového substantiva v současné češtině: sonda do korpusu. In Slovo a slovenost
66. RIV/zatím nebylo přiděleno
Štícha, František (2004): Sekundární imperfektiva v současné češtině. In Život s morfémy. Sborník studií na počest Zdenky Rusínové, pp. 151-160.
67. RIV/zatím nebylo přiděleno
Štícha, František (2004): Thematisierung, Satzanfang und Grammatikalität. In Linguistica Pragensia, pp. 90-103.
68. RIV/zatím nebylo přiděleno
Uhlířová, Ludmila (2004): Gramatika v korpusu, korpus v gramatice (Příspěvek k diskusi o vyhledávání gramatické informace v korpusech). In Slovo a slovenost
69. RIV/zatím nebylo přiděleno
Uhlířová, Ludmila (2004): O „nepřesné“ anafoře. In in press
70. RIV/68378092:_____/03:38030072
Uhlířová, Ludmila (2004): Samostatný lexém tento jako prvek množiny odkazových konkurentů. In Sborník prací Filozoficko-přírodovědecké fakulty Slezské univerzity v Opavě, pp. 168-176.
71. RIV/zatím nebylo přiděleno
Urešová, Zdeňka (2004): The verbal valency in the Prague Dependency Treebank from the annotator's point of view. In sborník přednášek JÚLŠ SAV (in press)
72. RIV/zatím nebylo přiděleno
Veselá, Kateřina; Havelka, Jiří; Hajičová, Eva (2004): Annotators' Agreement: The Case of Topic-Focus Articulation. In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 2191-2194. European Language Resources Association.
73. RIV/zatím nebylo přiděleno
Veselá, Kateřina; Havelka, Jiří; Hajičová, Eva (2004): Condition of Projectivity in the Underlying Dependency Structures. In Proceedings of Coling 2004, pp. 289--295. COLING.
74. RIV/zatím nebylo přiděleno
Veselá, Kateřina; Peterek, Nino; Hajičová, Eva (2004): Prosodic Characteristics of Czech Contrastive Topic. In Proceedings of 8th International Conference on Spoken Language Processing, Interspeech 2004, pp. 4. Sunjin Printing Co..
75. RIV/zatím nebylo přiděleno
Zeman, Daniel (2004): Data-Oriented Parsing by Rens Bod, Remko Scha, and Khalil Sima. In Prague Bulletin of Mathematical Linguistics, vol. 81 Univerzita Karlova.
76. RIV/zatím nebylo přiděleno
Zeman, Daniel (2004): Non-projectivity in Czech sentences. In UFAL/CKL technical report MFF UK, TR-2004-22.

77. RIV/zatím nebylo přiděleno
Zeman, Daniel (2004): Parsing with a Statistical Dependency Model. Univerzita Karlova.
78. RIV/zatím nebylo přiděleno
Žabokrtský, Zdeněk; Lopatková, Markéta (2004): Valency Frames of Czech Verbs in VALLEX 1.0. In HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation, pp. 70--77. Association for Computational Linguistics.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Specifikace a zdůvodnění jednotlivých položek finančních prostředků projektu čerpaných v r. 2004

1. rozpis celkových výdajů ve sledovaném období za všechny účastníky projektu

Prostředky	**Centrum**
Investice	**3 500**
Neinvestice	**31 943**
Mzdy	11 576
z toho OON	600
Režie	7 909
Ostatní	12 458
Ostatní podrobně	
Odpisy	2 839
Cestovné a pobyty	3 048
Pojištění	4 051
DHM a NHM	843
Materiál	539
Služby	249
Další	889
Celkem	**35 443**

2. specifikace a zdůvodnění jednotlivých výdajových položek ve vztahu k projektu

V roce 2004 byly všechny položky rozpočtu čerpány podle specifikace ve smlouvě. Zde uvádíme zdůvodnění podle jednotlivých pracovišť Centra.

Pracoviště MFF UK Praha:

Investice viz bod 3.

Neinvestice (plán 19 066 tis., využito 19 066 tis. Kč)

Mzdy (plán 9 983 tis.)
Mzdové prostředky podle plánu využity na platy a odměny zaměstnanců, viz bod 2. Personální a organizační zabezpečení Centra (využito 9 983 tis.).

OON (plán 600 tis.)
Prostředky využity podle plánu (využito 595 tis.).

Režie (plán 1 514 tis.)
Prostředky na režii využity podle plánu (využito 1 513 tis).

Ostatní (plán 7 569 tis., využito 7 570 tis)

Cestovné a pobyty (plán 2 638 tis.)
Cestovné bylo podle plánu použito na úhradu cest pracovníků CKL na zahraniční i domácí konference, kde prezentovali své výsledky (viz seznam zahraničních cest níže)

(využito 2 663 tis. Kč, mírné navýšení bylo hrazeno z nevyčerpaných prostředků na pojištění OON).

Pojištění (plán 3 494 tis.)
Prostředky využity podle plánu (využito 3 470 tis. Kč).

DHM a NHM (plán 586 tis.)
Prostředky využity podle plánu (využito 584 tis. Kč).

Materiál (plán 457 tis.)
Prostředky využity podle plánu (využito 458 tis. Kč).

Služby (plán 74 tis.)
Prostředky využity podle plánu (využito 74 tis. Kč).

Další (plán 320 tis.)
Prostředky využity podle plánu (využito 321 tis. Kč).

Seznam domácích a zahraničních cest pracovníků MFF UK hrazených částečně nebo zcela z prostředků CKL:

Pracovníci Centra věku většího 35 let

Jan Hajič
- **Brno, ČR, únor 2004**
  Účast na konferenci.

Eva Hajičová
- **Berlín, Německo, leden 2004**
  Zasedání Konsorcia projektu ERA (EU) Lang-Net, kde CKL/ÚFAL je navrhován za jednoho z partnerů.
- **Tübingen, Německo, leden - únor 2004**
  a) Účast na konferenci: Linguisticc Evidence-Empircal, Theoretical, and Computational Perspectives (Tübingen)
  b) Účast na konferenci: "Information Structure and the Architecture of Grammar (předsedání zasedání).
  C) přednáška na univ. Ve Stuttgartu.
- **Osaka, Japonsko, březen 2004**
  Účast na zasedání výkonného výboru International Speech & Communnication Association v Kyotu (jako členka výboru , (20. 22. 3. 2004)
  Účast na mezinárodní konferenci Speech Prosody 2004v Naře, 23. -26. 3. 2004.
- **Basilej, Švýcarsko, duben 2004**
  Účast na mezinárodním workshopu Metodologické základy studia mrtvých jazyků, přednesení pozvané přednášky.
- **Boston, USA, květen 2004**
  1. Účast na mezinárodní konferenci HLT/ACL 2004 v Bostonu. Proslovení referátu.
  2. Konzultace na Columbia University (Computer Science s prof. Kathy McKeown.
- **Lisabon, Portugalsko, květen 2004**
  Účast na konferenci LREC 2004, přednesení referátu společně s K. Veselou a J. Havelkou,
  2. Spoluorganizátorka 2 workshopů,
  3. Účast na zasedání přípravného výboru mezinárodního projektu EU ERA-Net.
- **Lipsko, Německo, červen 2004**
  Účast na mezinárodním workshopu Interface and Interface Conditions s přednesením referátu (společně s P. Sgallem) "Contextual Boundness and Context.
- **Paříž, Francie, červenec 2004**
  Účast na pracovní poradě partnerů společného navrhovaného projektu EU LangNet (v rámci programu ERA), příprava formulací, etap a předpokládaných výsledků.
- **Barcelona, Španělsko, červenec 2004**
  1. Účast na koordinačním zasedání k přípravě projektu LANGNET (program EU) - 19. 7.
  2. Účast na koordinačním zasedání pro přípravu tutorialů a značkování korpusu - 21. 7.
  3. Tutorial (45 účastníků) o anotování korpusu
  4. Hlavní konference ACL 2004 22. -24. 7.
  5. Členka programového výboru a účast na workshopu o discursu.l
- **Ženeva, Švýcarsko, srpen 2004**
  1.Účast na mezinárodní konferenci COLING 2004 přednsení referátu
  2. Jako místopředsedkyně mez. Komitétu komputační lingvistiky, účast na zasedání komitétu
  3. Účast na workshopu Dependency Grammer
- **Brno, ČR, září 2004**
  Návštěva MU Brno spolu s doc. Kučerou.
- **Leiden, Holandsko, září 2004**
  Účast na zasedání výkonného výboru - Commité International Permanent des Linguistes (jehož jsem členkou).
- **Azorské ostrovy, Portugalsko, září 2004**
  Účast na mezinárodním kolokviu DAARC 2004 (5th Discourse Anaphora and Anaphor Resolution Colloquium(ve

Furnas, (S. Miguel, Azores) ve dnech 23.-24. 9. 2004 s přednesením referátu Coreferential relations in the Prague Dependency Treebank (společný referát s Lucií Kučovou.

- Jejn Island, Jižní Korea, říjen 2004

1. Účast na zasedání výkonného výboru ISCA (Mezinárodní asociace pro řeč a jazyk), jehož jsem členkou (2. - 4. 10. 2004),
2. Účast na Int. Conf. Of Speech and Language Processing, Jejn Island, přednesení posteru (společný s N. Peterkem a K. Veselou) "Prosodic Characterics of Czech contrastive topic" (4.-7. 10. 2004),
3. Pozvaná přednáška na Seoul National University "Topic-Focus Articulation in the underlying Structure of Sentens" (8. 10. 2004).

- Šlapanice, ČR, listopad 2004

Účast na semináři.

- Paříž, Francie, listopad 2004

Jednání s vedením ústavu počítačové lingvistiky INALCO o zapojení do společenských projektů EU a do mezinárodní dohody.
Přednáška pro studenty a postgraduanty.

Markéta Lopatková

- Lisabon, Portugalsko, květen 2004

Účast na mezinárodní konferenci LREC 2004.
Účast na workshopu Building Lexical Resources.

- Brno, ČR, září 2004

Účast na konferenci TSD.

- Popradské Pleso, Slovensko, září 2004

Účast na mezinárodním workshopu ITAT 2004, přednesení příspěvku Dependency Analysis by Reduation for Natural Languages.

Jarmila Panevová

- Marburg, Německo, říjen 2004

Účast na odborném a organizačním zasedání Komise programtickou stavbu slovanských jazyků při Mezinárodním komitétu slavistů. Přednesení příspěvku, řízení organizačního zasedání (jako předsedkyně této komise).

- Olomouc, ČR, říjen - listopad 2004

Účast na semináři.

- Šlapanice, ČR, listopad 2004

Účast na semináři.

- Paříž, Francie, listopad 2004

Spolupráce se skupinou automatického zpracování jazyka v INALCO Paříž. Výuka pro studenty magistreského a doktorksého studia. Konzultace o spoluráci v v rámci projektů EU (M.-A. Moreaux, M. Faunton, M. Slasian).

Petr Sgall

- Tübingen, Německo, leden – únor 2004

Účast na dvou konferencích na univerzitě v Tübingen
9Linguistic Evidenc-Empirical, Theoretical and Computational Perspective) jako předseda zasedání , a Information Structure and the Architecture of Grammar , účast na panelu

- Brno, ČR, únor 2004

Účast na semináři.

- Lisabon, Portugalsko, květen 2004

Účast na mezinárodní konferenci LREC 2004. Účast na jednání přípravného výboru mezinárodního projektu ERA-Net, účast na workshopu o lexikální sémantice.

- Lipsko, Německo, červen 2004

Aktivní účast na konferenci "Interface and Interface Conditions s přednesením referátu, společně s E. Hajičovou přednesení referátu "Contextual Boundness and Context.

Zdeňka Urešová

- Philadelphia, Baltimore, USA, červen - červenec 2004

1. Pobyt na University of Pennsylvania, PA = příprava tutoriálu na ACL 04, Barcelona (20.-22. 7. 04) demonstarce (pro CIS a LDC) syntaktické anotace PDT, jednání o publikaci CD v LDC.
2. Pobyt u IBM Research Yorktown, projekt MALACH, jednání o přídavném projektu předkládaném NSF USA.

- Barcelona, Španělsko, červenec 2004

Účast na tutoriálu (spoluautorka) "Beyond Syntax: Valency, Predicate - argument Structure and FrameNet", 21. 7. 04, 45 účastníků
- Účast na schůzce o projektech sémantické anotace, 21. 7. 2004
- Účast na ACL 2004, 22. 7. - 24. 7. , účast na workshopu Text and Meaning,
- Demonstrace syntaktické anotace na AR a TR, demonstrac PDT - Vallexu na University of the Basque Country, San Sebastina, 28.- 29. 7. 2004

Pracovníci Centra věku menšího či rovna 35 let

Emanuel Beška

- Káhira, Egypt, září 2004

Podíl na prezentaci v rámci konference NEMLAR v Káhiře. Šlo o příspěvek Prague Arabic Dependency Treebank - Development in Data and Tools, zahrnující výsledky dotyčného projektu na ÚFAL a CKL UK MFF.

Ondřej Bojar

- Nancy, Ženeva, Francie, Švýcarsko, srpen 2004

V účast na letní škole ESSLLI. V Ženevě účast na konferenci COLING 2004 a workshopu věnovaného závislostním gramatikám.

- Kodaň, Dánsko, září 2004

Na workshopu Constraint Solving and Language Processing (CSLP 2044) přednesení referátu s názvem "Problems of Large Coverage Constrait-Basel Dependency Grammer for Czech.

Silvie Cinková

- Brno, ČR, září 2004

Účast na konferenci TSD.

Jan Cuřín

- Boston, USA, květen 2004

Účast na konferenci HLT/NAACL/04
Účast na workshopu Frontiers in Corppus Annotation"
Spoluaotor příspěvku "Prague Czech-English Dependency Treebank: Any hopes for common annotation scheme_"
Návštěva na Brown University v Providence (E. Charniak, K. Hall, H. Fox), jednání o možnosti vužití závislostního přístupu ve strojovém překladu

- Lisabon, Portugalsko, květen 2004

Účast na konferenci LREC 2004, spoluautor příspěvku s názvem:
"Prague Czech-English Dependency Treebank: Syntactically Anotated Resources for Structual Machine Translation"

Martin Čmejrek

- Boston, USA, květen 2004

1. Účast na konferenci HLT/ACL 2004 proslovení referátu
2. Návštěva na Brown University v Providence. Jednání o problematice strojového překladu s E. Charniakem, H. Fox a K. Hall.

- Lisabon, Portugalsko, květen 2004

Účast na konferenci LREC 2004, spoluautor příspěvku s názvem:
"Prague Czech-English Dependency Treebank: Syntactically Anotated Resources for Structual Machine Translation"

Martin Holub

- Brno, ČR, leden – únor 2004

Účast na semináři

- Barcelona, Španělsko, červenec 2004

Aktivní účast na konferenci ACL 2004,
publikovaný příspěvek ve sborníku, přednesený referát.

Petr Homola

- Riga, Litva, duben 2004

Účast na konferenci a přednesení referátu.

- Saarbrücken, Německo, duben 2004

1) na DFKI (s J. Piskorskim a W. Drozdynskym) práce na integraci zdrojů pro slovanské jazyky do systému SProllT a s tím spojeným vylepšením s jazyky Česílko (pro pár CZ-PL)
2) na CeLi (G.-J. Kruijff) práce na statistických jazykových modelech (včetně pro češtinu)

- Zakopane, Polsko, květen 2004

Prezentace článku o využití Named entity recognition v systému strojového překladu.

- Kaunas, Litva, červen 2004

Účast na konferenci "KTU - 3rd international conference Language, technology and culture variety" , přednesení referátu "Machine translation among related languages". Vypracování konceptu systému strojového překladu pro litevštinu založeného na hloubkové analýze..

- Nancy, Francie, srpen 2004

Na letní škole účast na několika týdenních kurzech, na 'student session' přednesení referátu na téma "On some aspects of machine translation among related languages".

Veronika Kolářová

- Brno, ČR, září 2004

Účast na konferenc TSD.

- Šlapanice, ČR, listopad 2004

Lucie Kučová
- **Azorské ostrovy, Portugalsko, září 2004**
  Účast na konferenci DAARC 2004 jako spoluautor příspěvku, jenž byl na této konferenci pronesen.

Pavel Pecina
- **Lisabon, Portugalsko, květen 2004**
  Účast na konferenci LREC 2004. Prezentace referátu na workshopu Building Lexical Resources..

Petr Podveský
- **Nancy, Francie, srpen 2004**
  ESSLLI 2004, letní škola o logice jazyka a jejich strojovém zpracování, účast na kurzech a na akcích spojených s letní školou.

Kiril Ribarov
- **Lisabon, Portugalsko, květen 2004**
  Účast na mezinárodní konferenci LREC 2004 a na seminářích během konference. Aktivní účast v sekci "Tools for Corpora 7 Lexicons", kde jsem prezentoval svoji práci na anotačním nástroji ACT pro psané kulturní dědictví.

Jiří Semecký
- **Barcelona, Španělsko, červenec 2004**
  Účast na konferenci ACL 2004. Spolu s Martinem Holubem a Jiřím Divišem na studentské sekci (Student Research Workshop) při ACL -04 publikace článku Seaching for Topics in a Large Collection of Texts.

Otakar Smrž
- **Káhira, Egypt, září 2004**
  Prezentace výsledků projektu Prague Arabic Dependency treebank ve fromě dvou konferenčních příspěvků

Pavel Straňák
- **Brno, ČR, leden 2004**
  Účast na semináři.
- **Berlín, Německo, únor 2004**
  Konzultace a jednání s prof. Hanksem a Annou Rumschiski.
- **Lisabon, Portugalsko, květen 2004**
  Účast na konferenci LREC 2004.

Kateřina Veselá
- **Lisabon, Portugalsko, květen 2004**
  Účast na konferenci LREC 2004 - sekce Coreference and Anapohora, Tagging, Evaluation a účast na workshopu. Spoluautor referátu
- **Ženeva, Švýcarsko, srpen 2004**
  Účast na konferecni COLING 2004, referát "Condition of Projectivity in the Underlying Dependency Structures (společně s E. Hajičovou aj. Havelkou).

Daniel Zeman
- **Ženeva, Švýcarsko, srpen 2004**
  Účast na mezinárodní vědecké konferenci COLING 2004, pořádané Ženevskou univerzitou, včetně přidružených workshopů.
- **Brno, ČR, září 2004**
  Účast na konferenci TSD.

Zdeněk Žabokrtský
- **Boston, USA, květen 2004**
  Účast na konferenci Human Language Technology 2004 a na workshopu Forntiiers in Corpus Annotation přednesení příspěvku Valency Frames of Czech Verbs in VALLEX 1.0. Návštěva prof. Charniaka na Brown University.
- **Saarbrücken, Německo, červenec 2004**
  Prezentace valenčního slovníku VALEX, seznámení se s jejich přístupem k lexikální sémantice v projektu SALSA. Práce s tektogramatickými stromy vytvořenými anotací části korpusu Nogra, práce na zlepšení lematizace vstupních databází a odstraňování nekonzistencí v anotovaných datech a zprovoznění nových nástrojů pro práci s daty.

Pracoviště ZČU Plzeň:

Investice: Celková cena 400 tis. Kč, detailní přehled viz bod 2.

Neinvestice (přiděleno 1 839 tis. Kč)

Mzdy (plán 603 tis. Kč)

Za měsíce leden až prosinec 2004 bylo vyplaceno		**603 tis. Kč**

OON (plán 0 Kč)

Za měsíce leden až prosinec 2004 bylo vyplaceno		**0 tis. Kč**

Režie (plán 206 tis. Kč)

Režie		**206 tis. Kč**

Ostatní (plán 1 030 tis. Kč)

Cestovné (plán 234 tis. Kč)

Konference ICASSP, ICSLP, TSD, SPECOM,		144,2 tis. Kč
Meeting týmu letního workshopu (pořádala JHU Baltimore)		38,8 tis. Kč
Workshop CLEF 2004 v Bathu		15,0 tis. Kč
Setkání týmu IEM v Londýně		16,6 tis. Kč
Cestovné tuzemské		25,8 tis. Kč
	Celkem	**240,4 tis. Kč**

Pojištění sociální a zdravotní (plán 211 tis. Kč)

Pojištění (leden-prosinec)		**210,1 tis. Kč**

DHM a NHM (plán 110 tis. Kč)

Přenosné disky		18,1 tis. Kč
2x HDD WD		25,7 tis. Kč
Sluchátka		3,5 tis. Kč
Knihy		42,4 tis. Kč
Řečové korpusy (LDC)		26,3 tis. Kč
Propojovací kabely		2,2 tis. Kč
	Celkem	**118,2 tis. Kč**

Materiál (plán 35 tis. Kč)

Kancelářský materiál		12,5 tis. Kč
MMVS kupony		1,5 tis. Kč
Pořadač CD		1,6 tis. Kč
	Celkem	**15,6 tis. Kč**

Služby (plán 146 tis. Kč)

Vložné TSD 2004		10,8 tis. Kč
Vložné na konf. (ICASSP2004, SPECOM2004, ICSLP2004, CLEF2004)		57,6 tis. Kč
Publikační služby		70,0 tis. Kč
Ostatní služby (bankovní, telefonní, poštovné, přepravné ap.)		12,5 tis. Kč
	Celkem	**150,9 tis. Kč**

Další (plán 294 tis. Kč)

Místnosti (teplo, energie ap.)		100,0 tis. Kč
Stipendia		160,0 tis. Kč
Oprava a údržba		29,3 tis. Kč
Kopírovací služby		5,5 tis. Kč
	Celkem	**294,8 tis. Kč**

Neinvestice celkově (k 11.1.05) **1 839,0 Kč**

Seznam cest pracovníků ZČU hrazených částečně nebo zcela z prostředků CKL:

- Na prestižní konferenci ICASSP v Montrealu byl přednesen vyzvaný referát: Psutka, J., Hajič, J., Byrne, W.: The Development of ASR for Slavic Languages in the MALACH Project (prof. Psutka), na konferenci SPECOM v Petrohradě byl presentován článek: Kolář, J., Švec, J., Psutka, J.: Automatic Punctuation Annotation in Czecch Broadcast News Speech (ing. Kolář, doktorand podílející se na práci CKL). Na konferenci TSD - prof. Psutka chairman sekce „Speech processing".
- Doktorand spolupracující s CKL Ing. Kolář se zúčastnil konference Conference SPECOM2004 (Petersburg).
- Pracovník Centra Ph.D. Pavel Ircing byl vybrán jako člen pracovního týmu pro letní Workshop pořádaný JHU v Baltimore. Tématika řešená na workshopu velmi úzce souvisela se zaměřením nově navrhovaného „Centra komputační a aplikované lingvistiky" (pokračovatel CKL), jehož měl být Ph.D. Ircing klíčový pracovník. Z projektu byly placeny jen částečně náklady na dvě pracovní setkání (první na JHU v Baltimore a druhé souběžně s konferencí ICASSP v Montrealu)
- Ph.D. Pavel Ircing se zúčastnil workshopu CLEF2004 (Cross-Language Evaluation Forum) v Bathu. CLEF se zabývá výzkumem metod pro „information retrieval" v multijazykovém prostředí. P. Ircing byl přizván do pracovní skupiny, kde projednával možnosti zařazení CKL a zejména navrhovaného pokračujícího Centra do těchto mezinárodních aktivit (navrhovaná tématika velmi úzce souvisí s tématikou plánovaného pokračujícího „Centra komputační a aplikované lingvistiky").
- prof. Psutka se zúčastnil v Londýně pracovní schůzky navrhovatelů projektu „Immersive Environment Machines" v rámci 6. RP EU. Plzeňská sekce Centra by měla na tomto velmi rozsáhlém projektu řešit problematiku zpracování informací získaných z „řečového kanálu". Opět šlo o tématiku, se kterou bylo počítáno v navrhovaném pokračujícím „Centru komputační a aplikované lingvistiky".

Pracoviště UJČ Praha:

Investice: Celková cena 200 tis. Kč, detailní přehled viz bod 2.

Neinvestice: Neinvestice jsou čerpány podle rozpočtu na rok 2004.

Seznam cest pracovníků UJČ hrazených částečně nebo zcela z prostředků CKL:

Pracovníci Centra věku většího 35 let

F. Štícha:
březen 2004: účast na konferenci v Mannheimu
květen 2004: pracovní pobyt na univerzitě v Neapoli (spolupráce s prof. F. Esvanem na tvorbě vidových databází)
září 2004: pracovní seminář na univerzitě v Tübingen (spolupráce s prof. T. Bergerem na získávání dat o české morfologii z korpusů)
září 2004: účast na konferenci o anafoře a korpusech na Azorských ostrovech
listopad 2004: účast na konferenci v Salzburgu (referát na téma Grammar and Corpus)
listopad 2004: účast na konferenci v Regensburgu o modalitě ve slovanských jazycích (korpusový referát)

L. Uhlířová:
říjen 2004, Marburg, Německu, zasedání mezinárodní komise pro gramatickou stavbu slovanských jazyků;
Sofia, duben, Sofijská univerzita; Ústavu bulharského jazyka

Pracovníci Centra věku menšího 35 let

Flanderková, E.:
11. - 18. 11. 2004, studijní pobyt v Max-Planck-Institut for Psycholinguistics, Holandsko, Nijmegen

Prošek, M.:
8. 9. - 11. 9. 2004 Slovensko, Bratislava, "Konferencia o jazykovej kultúre", konaná Jazykovedným ústavem L. Štúra SAV.

Pravdová, M., Smejkalová, K., Prošek M.:
22. 9. - 27. 9. 2004 Itálie, Bergamo, slavistická konference "Polyslav"

3. jednoznačná specifikace položek hrazených z účelové dotace

Plán:

Účelové prostředky	Centrum	MFF UK	ZČU	ÚJČ
Investice	**3 500**	**2 900**	**400**	**200**
Neinvestice	**22 682**	**19 066**	**1 839**	**1 777**
Mzdy	11 318	9 983	603	732
z toho OON	600	600	0	0
Režie	1 894	1 514	206	174
Ostatní	9 470	7 569	1 030	871
Ostatní podrobně				
Cestovné a pobyty	3 048	2 638	234	176
Pojištění	3 961	3 494	211	256
DHM a NHM	843	586	111	146
Materiál	539	457	35	47
Služby	249	74	146	29
Další	830	320	293	217
Celkem	**26 182**	**21 966**	**2 239**	**1 977**

Skutečnost:

Účelové prostředky	Centrum	MFF UK	ZČU	ÚJČ
Investice	**3 500**	**2 900**	**400**	**200**
Neinvestice	**22 682**	**19 066**	**1 839**	**1 777**
Mzdy	11 318	9 983	603	732
z toho OON	595	595	0	0
Režie	1 893	1 513	206	174
Ostatní	9 471	7 570	1 030	871
Ostatní podrobně				
Cestovné a pobyty	3 085	2 663	240	182
Pojištění	3 946	3 470	210	266
DHM a NHM	862	584	118	160
Materiál	519	458	16	45
Služby	254	74	151	29
Další	805	321	295	189
Celkem	**26 182**	**21 966**	**2 239**	**1 977**

4. specifikace položek hrazených z prostředků příjemce, příp. spolupříjemců.

Prostředky nositele	**Celkem**	**MFF UK**	**ZČU**	**ÚJČ**
Investice	**0**	**0**	**0**	**0**
Neinvestice	**9 261**	**7 840**	**489**	**932**
Mzdy	258	172	0	86
Režie	6 015	5 113	221	681
Odpisy	2 838	2 475	239	125
Místnosti	20	0	20	0
Pojištění	90	60	0	30
Další	40	20	10	10
Celkem	**9 261**	**7 840**	**489**	**932**

Komentář: Vklad ZČU byl realizován na účet projektu ve výši 251 tis. Kč. Tento vklad byl využit na částečné pokrytí režijních nákladů Centra. Další spoluúčast ZČU byla provedena prostřednictvím odpisů investičního majetku. ZČU hradí odpisy investičního majetku zakoupeného v plzeňské sekci CKL v roce 2000, 2002, 2003 a v roce 2004 (v roce 2001 byly investice nakupovány MFF a jsou i v jejím majetku). Bohužel není administrativně možné, aby odpisy majetku CKL realizované na ZČU procházely účetně přes zvláštní účet otevřený pro vykazování spoluúčasti ZČU.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Přehled čerpání finančních prostředků projektu v době řešení projektu

Náklady na řešení projektu

Rok	**Účelová podpora ze státního rozpočtu (tis. Kč)**	**Jiné zdroje použité k řešení projektu (tis. Kč)**	**Typy jiných zdrojů** (veřejné jiné než účelová. podpora, tuzemské neveřejné, zahraniční atp.)
2000	10 662	3 177	veřejné jiné než účelová
2001	18 671	6 387	veřejné jiné než účelová
2002	21 860	7 764	veřejné jiné než účelová
2003	20 520	8 666	veřejné jiné než účelová
2004	26 182	9 261	veřejné jiné než účelová
Celkem	97 895	35 255	

Zpracoval (jméno): Ing. Vlad. Stáňa, vedoucí HO

V dne

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208

Tisková zpráva [2]***:***

Cílem Centra komputační lingvistiky (CKL) byl výzkum a vývoj v oblasti moderní počítačové lingvistiky na zcela nové úrovni založené na jedinečné vícerovinné analýze velmi rozsáhlého korpusu. Činnost Centra, díky kterému se podařilo vytvořit u nás jediný integrovaný tým pro výzkum psané i mluvené řeči, měla a má velký význam pro aplikace v mnoha oborech služeb a průmyslu, které pracují s komunikací člověka s počítačem.

Stěžejním projektem CKL bylo vybudování Pražského závislostního korpusu, což je soubor českých textů s bohatou informací o morfologii, větné stavbě a významové struktuře vět (první verze korpusu, "Prague Dependency Treebank, Version 1.0" byla vydána na CD-ROM v roce 2001, druhá verze, "Prague Dependency Treebank, Version 2.0" bude vydána v roce 2005). Takový soubor textů slouží jednak dalšímu teoretickému zkoumání češtiny, zejména jde však o velké množství lingvisticky zpracovaných dat, která jsou nezbytná pro automatické zpracování přirozeného jazyka pro jakýkoliv aplikovaný úkol – strojový překlad, vyhledávání informací (tzv. data mining), automatické "porozumění" textu i jeho generování.

Druhým základním směrem Centra byl statisticky založený výzkum v oblasti rozeznávání mluvené řeči. Výsledky tohoto směr výzkumu byly dán k dispozici odborné veřejnosti jako "Czech Broadcast News Corpus" a " Czech Broadcast News Transcripts" na dvou CD-ROM v roce 2004. Zásadním přínosem bylo zapojení Centra do mimořádně rozsáhlého mezinárodního projektu MALACH (Multilingual Access to Large Spoken Archives), jehož cílem je vývoj systémů pro automatický předpis svědeckých výpovědí lidí, kteří přežili holocaust. Svědecké výpovědi byly pořízeny ve více než 30 různých jazycích a česká strana je prostřednictvím Centra spoluzodpovědná za zpracování jazyků střední a východní Evropy.

Dalším cílem výzkumu Centra bylo vytváření a využívání vícejazyčných zdrojů. Pozornost byla věnována zejména studiu a uplatnění paralelních korpusů se zaměřením na strojový a strojem podporovaný překlad – v roce 2004 byla vydána unikátní sada počítačových databází a nástrojů pro automatický překlad "Prague Czech-English Dependency Treebank, PCEDT 1.0". Takto pojatá výzkumná činnost vedla k získání dalších znalostí o češtině srovnatelných s výsledky výzkumu jiných jazyků.

Nepostradatelnou součástí činnosti Centra komputační lingvistiky jako centra základního výzkumu byl výzkum teoretických aspektů komputační lingvistiky se zaměřením především na češtinu v podobě psané i mluvené a s ohledem na možné aplikace. Metodologie výzkumu v rámci Centra byla založena na prohloubeném studiu, porovnávání a kvalifikovaném využití postupů strukturních i statistických včetně metod strojového učení, s ohledem na specifické typologické vlastnosti češtiny jako vysoce flexívního jazyka.

Jak ukázala veřejná vědecká rozprava o výsledcích Centra konaná ve dnech 29.-30. listopadu 2004, za účasti 7 předních zahraničních vědců z oboru komputační lingvistiky, tyto výsledky mají přední místo v evropském i světovém výzkumu a jsou ve světě přijímány s vynikajícím ohlasem.

Činnost Centra bohatě naplnila očekávané možnosti v navazování a udržování těsných kontaktů s českým a mezinárodním průmyslem využívajícím počítače, o čemž svědčí i zájem partnerů a uživatelů z oblasti aplikační sféry o vhodně zpracované a užitečné zdroje pro široce založený vývoj a aplikace.

V Praze dne:

řešitel projektu
(podpis)

příjemce dotace
(razítko a podpis statut .zást. nositele)

[2] Tisková zpráva je součástí pouze závěrečné zprávy a charakterizuje hlavní dosažené výsledky projektu, (záznamy o konkrétních výstupech projektu jako jsou publikace, výzkumné zprávy, patenty atd. nositel zasílá každoročně do RIV!¨.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace::	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1, IČO: 00216208