Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1,110 00, IČO: 00216208

Průběžná zpráva o realizaci projektu[1] v roce 2003

1. Stručný přehled dílčích cílů projektu splněných v uplynulém období

Uvádíme zde výsledky výzkumné činnosti v rámci projektu rozčleněné podle jednotlivých bodů programu pro rok 2003 (T1 až T6, viz Přehled a upřesnění dílčích cílů projektu a postupu při jejich naplňování pro rok 2003). Cíle jsou konkretizovány v rámci tří nosných výzkumných projektů, a to rozvíjení Pražského závislostního korpusu (bod B původního návrhu, zde T-1), projekt strojového překladu (bod F původního návrhu, zde T-2) a v rámci výzkumu v oblasti zpracování mluvené řeči pak participace na mimořádně rozsáhlém mezinárodním projektu MALACH (bod E původního návrhu, zde T-3). Souběžně s těmito projekty a v návaznosti na ně pokračoval výzkum v oblasti teoretických aspektů komputační lingvistiky, tedy jejích matematických i lingvistických základů (body A, C a D, zde T-4) a rovněž vyvíjení některých aplikačních systémů (bod F původního návrhu, zde T-5).
V následujícím přehledu jsou jednotlivé body konkretizovaného programu pro rok 2003 označeny (T1) až (T6), popř. dalším členěním podle dílčích cílů plánovaných na rok 2003 (v souladu s časovým harmonogramem). Věcnou část zprávy doplňujeme seznamem dat a souvisejících nástrojů získaných v rámci realizace projektu (Příloha 1) a seznamem publikací v r. 2003 a prací připravených k publikaci (Příloha 2).

T1: Rozvíjení Pražského závislostního korpusu (bod B původního návrhu)

Pražský závislostní korpus (Prague Dependency Treebank, PDT) byl nadále stěžejním projektem CKL – výzkum v roce 2003 plynule navazoval na výsledky let předcházejících. Pozornost byla zaměřena hlavně na dokončení anotace 55 000 vět (1 050 souborů po 50 větách) na tektogramatické rovině a zahájena kontrola jejich konzistence tak, aby v následujícím roce bylo možno dát celý zkontrolovaný anotovaný soubor o délce 800 000 slov k dispozici na CD-ROM pro širší odbornou veřejnost a také pro využití zájemci z nejrůznějších oblastí aplikační činnosti. Výsledkem je vůbec první takto široce pojatý anotovaný korpus přirozeného jazyka na podkladové rovině, jak o tom svědčí výrazně kladné zahraniční ohlasy a zájem o koncepci a metodu práce na projektu.

T1-1 V oblasti jazykové databáze se pracuje na doplňování a modifikacích morfologického slovníku češtiny. Byla vytvořena pracovní verze specifikace slovníku pomocí technologie XML, obousměrně kompatibilní se specifikací původní. Dále byl morfologický slovník doplněn o názvy obcí ČR. Systém morfologických vzorů byl pozměněn, pracuje se na rozšíření systému značek, vše při zachování kompatibility se starší verzí.
Na úrovni hloubkové struktury byl dále rozšiřován valenční slovník PDT-VALLEX, který byl v z důvodů konzistence v minulém období integrován do anotačního schématu tektogramatické roviny (5 200 položek slovníku, 8 300 valenčních rámců). Byly vyhodnoceny rozdíly mezi anotátory a provedena systematická ruční kontrola jeho konzistence. Dále byla sjednocena anotace jednotlivých položek slovníku a chybějící položky byly doplněny. Odpovídající slovník pro substantiva bude následovat v prvních týdnech roku 2004.
Pokračovala práce na komplexně anotovaném valenčním slovníku sloves VALLEX, jehož první verze byla dána k dispozici odborné veřejnosti (1400 položek slovníku, 4 000 valenčních rámců, viz technická zpráva TR-2003-18).

T1-2 Byly zpřesňovány podmínky pro automatickou specifikaci valenčních rámců neslovesných slovních druhů. Vedle pravidelných změn povrchových realizací valenčních doplnění u syntakticky derivovaných substantiv bylo zkoumáno valenční chování substantiv se zabudovanou rolí; u činitelských jmen (zabudovaný ACT, např. dodavatel) a u substantiv se

[1] Zpráva podepsaná řešitelem, která byla schválena oponentním řízením, se současně se zápisem o oponentním řízení, (pokud bylo pořádáno) vyúčtováním za uplynulé období, upřesněním dílčích cílů a rozpočtu pro následující období zasílá v jednom vyhotovení zadavateli, (závěrečná zpráva se zasílá ve dvou vyhotoveních).

zabudovaným PAT (např. dárek, výplata) je třeba počítat se změnou formy Dat → Gen (dodávat firmě → dodavatel firmy). Algoritmus pro substantiva odvozená lexikální derivací a pro substantiva odvozená syntaktickou derivací se tedy bude lišit.

Dále byla zkoumána deverbativní substantiva, která jsou součástí frazeologických slovesných sousloví (tato substantiva jsou při anotacích označena speciálním funktorem) a o deverbativní substantiva v plně substantivním užití.

T1-3 Byla upřesněna pravidla pro celkovou anotaci aktuálního členění, a tato pravidla byla včleněna do příslušné kapitoly Manuálu pro tektogramatické značkování (viz technická zpráva TR-2003-20). Byly implementovány programové nástroje (makra pro anotační nástroj TrEd) pro předvyplňování atributu TFA a pro modifikaci struktury věty při anotování aktuálního členění, zejména prostředky pro práci s neprojektivními stromy. Rozdíly ve značkování anotátorů byly vyhodnocovány na paralelně anotovaných souborech.

Byly zkoumány otázky aktuálního členění věty v souvislosti s větnou prozodií, takže byly spojeny práce na textech z korpusu psaných projevů s prací na mluvené řeči.

T1-4 Byla dokončena typologie koreference a byly načrtnuty přesahy do dalších oblastí anotace velkého souboru (lemmata doplněných uzlů). V rámci evaluace výsledků značkování koreferenčních vztahů byla provedena paralelní anotace s velmi dobrými výsledky v oblasti mezianotátorské shody (94%). Pokračovalo se ve vytváření automatické procedury zachycení gramatické koreference. Výsledky studia koreferenčních vztahů a možnosti jejich anotace byly shrnuty v technické zprávě (viz technická zpráva TR-2003-19).

K 30.11.2003 bylo dokončeno anotování koreference na 15 000 větách.

T1-5 Byla dokončena první fáze anotace tzv. velkého souboru (55.000 vět) na tektogramatické rovině (tj. rovině hloubkové syntaxe).

T1-6 Byla zahájena kontrola anotace na první podúrovni (tzv. velký soubor) ověřující jednak formální (technickou) správnost tektogramaticky anotovaných dat, jednak soulad reálné anotace s pokyny Manuálu pro anotátory. Byl sestaven a dále doplňován seznam automatických a poloautomatických kontrol, které jsou postupně implementovány.

Byl vyvinut speciální nástroj umožňující rozložit výpočetní i paměťové nároky mezi libovolné množství výpočetních strojů - tímto způsobem byl čas potřebný k jednomu průchodu kontrolním programem zredukován z desítek minut na jednotky, maximálně desítky sekund.

Klíčovým bodem probíhajících oprav je sjednocení způsobu anotace doplnění u sloves a substantiv na základě valenčního slovníku PDT-VALLEX (viz bod T1-1).

T1-7 Byla provedena řada experimentů s nástrojem, který využívá valenční slovník pro doplnění uzlů elidovaných na povrchu věty a vyplnění některých jejich atributů. Byly do něj implementovány některé lingvisticky motivované vlastnosti, byl pokusně použit na datech ze slovníku VALLEX i na valenčním slovníku vytvářeném anotátory (PDT-VALLEX) a byly prozkoumány způsoby kooperace tohoto nástroje s nástrojem přiřazujícím tektogramatické funktory. Výsledkem těchto experimentů je znalost optimálního způsobu použití nástroje a jeho vyšší účinnost.

T1-8 Na základě vyhledávacího nástroje NETGRAPH byly provedeny první účelové vyhledávací sondy, zejména v anotacích analyzovaných na analytické rovině. Získaný materiál sloužil k různým lingvistickým studiím, popř. i předběžným závěrům, zčásti byl materiál doplněn i vyhledávkami v ČNK.

Na spoluřešitelském pracovišti (UJČ) byla činnost orientována především na možnosti vyhledávání syntaktických struktur zapsaných v podobě stromů v PDT:

- Byly zkoumány a ověřovány možnosti vyhledávání gramatických struktur v PDT a výsledky Byly konfrontovány s nálezy v Českém národním korpuse (ČNK).
- Byly testovány možnosti, jak vyhledávat některé specifické slovosledné jevy (v rámci nominálni skupiny).
- Byly vyhledávány struktury s reflexivním deagentivem v PDT.

Na spoluřešitelském pracovišti (UJČ) byly vytvořeny databáze pro ukládání nálezů z Českého národního korpusu a z PDT (přístupné účastníkům projektů GAČR). Byla dokončena revize větší části archivovaných dokumentů, t.j. cca 4000 dokumentů z dosavadního počtu 4350 dokumentů.

Při poradenské činnosti v oddělení jazykové poradny a gramatiky (UJČ) se využívalo nálezů z PDT (konzultace syntaktických dotazů – např. značení větných členů) a pracovalo se s

poradenskou databází (konzultace dotazů, spolupráce na jejím vytváření).

T1-9 V šedesátých a sedmdesátých letech století vznikal v Ústavu pro jazyk český AV, pod vedením dr. Marie Těšitelové, první anotovaný korpus češtiny (tzv. Český anotovaný korpus, ČAK). Anotace probíhaly na dvou rovinách: morfologické a syntakticko-analytické. Z celkového objemu 600 000 slov dvě třetiny představují texty psané a jedna třetina jsou transkripce mluvených projevů; textový materiál obsahuje texty publicistické, vědecké a administrativní. Ve své době sloužil jako podkladový materiál pro statistické (kvantitativní) zpracování češtiny. Tento korpus je nyní převáděn do formátu PDT, což umožní dále využívat řadu nástrojů navržených pro PDT. V současné době dokončujeme konverzi morfologického anotování.

T2: Strojový překlad (bod F původního návrhu)

V oblasti strojového překladu výzkum plynule navázal na výsledky dosažené v roce 2002.

T2-1 V návaznosti na spolupráci s Johns Hopkins University byla dokončena nová implementace trénování překladových modelů z letního workshopu v Baltimore z r. 2002, probíhá testování a ladění. Probíhá vývoj dekodéru.

T2-2 Moduly pro analýzu češtiny byly rozšířeny o možnost začlenění modernějšího parseru. Byla dokončena základní verze překladového systému využívající pravidlového modulu pro generování angličtiny z tektogramatické reprezentace. Dále proběhla integrace systému s jazykovým modelem postaveným nad rozsáhlým anglickým korpusem.

Byl dokončen tzv. baseline systém pro strojový překlad. Systém obsahuje: automatickou morfologickou analýzu českého vstupu, výběr ze dvou možností statistického parsování do analytické reprezentace, pravidla pro převod do tektogramatické reprezentace, mechanismus pro filtrování překladových slovníků, lexikální transfer využívající strukturního kontextu, pravidla pro generování výstupní anglické věty a jazykový model pro výběr optimální varianty překladu.

T2-3 Pokračovalo obohacování slovníků vhodných pro strojový překlad, angličtina-čeština a čeština-angličtina. Byl implementován modul pro výběr vhodného překladu na základě strukturního kontextu.

Byl dokončen systém pro automatické filtrování slovníků pro strojový překlad. V rámci experimentu s interpolací skóre překladového modelu při automatickém rozpoznávání řeči byl vygenerován překladový česko-anglický slovník forem. Byla zahájena příprava slovníku pro vydání v LDC.

T2-4 Do systému byl začleněn evaluační model metodiky IBM (BLEU) pro automatické vyhodnocování výsledků systému.

Byla zahájena příprava programových nástrojů pro česko-anglický strojový překlad, které budou vydány v LDC v roce 2004.

T3: Zpracování mluvené řeči (bod E původního návrhu)

T3-1 V roce 2003 pokračovaly práce především na systémech rozpoznávání spontánní řeči v rámci velkého 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 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 (www.clsp.jhu.edu/research/malach), práce na konstrukci systému rozpoznávání řeči byly pak prováděny částečně na půdě CKL.

– Do systému pro rozpoznávání spontánní češtiny byla implementována metoda adaptace na řečníka. Již při prvních experimentech s rozpoznáváním češtiny v projektu MALACH bylo zjištěno, že promluvy zpovídaných svědků holocaustu obsahují velké množství nespisovných slov. Proto byla provedena úprava slovníku, při které byla ke každému nespisovnému slovu přidána příslušná spisovná varianta. Výsledkem těchto úprav je nový slovník, který umožňuje generovat na výstupu dekodéru alternativně spisovnou či

nespisovnou variantu. Přidání odpovídajících vah k jednotlivým nespisovným variantám spisovného slova vyústilo ve slibné zlepšení úspěšnosti rozpoznávání (větší než 2%). Též jazykový model, připravený z psaných (spisovných) textů vybíraných například z knih a novinových článků, nyní lépe souhlasí s takto modifikovaným přepisem výpovědí a dále zlepšuje výslednou funkci rozpoznávače.

- Pokračovala manuální morfologická anotace trénovacího textového korpusu. Jde o velmi náročný a pracný proces, a proto v současné době ještě není zpracován dostatek promluv pro experimenty. Výsledná data budou použita pro natrénování jazykových modelů založených na morfologických značkách.
- Paralelně byl testován morfologický model využívající automatické morfologické analýzy psaného textu. Tento model vylepšil úspěšnost rozpoznávání o 6%. Úspěšně jsme do systému rozpoznávání řeči zapojili HMM tagger.
- Dále byla vytvořena první verze systému pro rozpoznávání spontánní ruštiny. Tento systém musel být vybudován od základu, tj. od návrhu fonetické abecedy a pravidel pro fonetickou transkripci, přes úpravu anotačního programu, až po návrh fonetických rozhodovacích stromů. Z literatury není známo, zda byl podobný systém pro rozpoznávání souvislé ruštiny na nějakém pracovišti dosud realizován. V současné době je dokončována anotace ruských řečových dat a probíhá také kontrola kvality přepisu. Předpokládáme, že v průběhu roku 2004 bude dobudována verze rozpoznávače souvislé spontánní ruštiny včetně propojení s jazykovým modelem, pro jehož natrénování byla v průběhu roku též sbírána data.

T3-2 V rámci studia prozodických opozic aktuálního členění byly označkovány různé druhy základů a ohnisek u 1000 vět. Pro tyto věty byly spočítány průběhy prozodických parametrů, které jsou podkladem ke studiu prozodických opozic.

Na základě vět s označkováním různých druhů základů a ohnisek byly provedeny první analýzy prozodických opozic. Tyto analýzy zatím podporují hypotézu opozic pro kontrastivní základ a ohnisko. Prozodickou charakteristikou vybraných úseků byla F0 křivka.

T3-3 Prozodická databáze je v současné době rozšiřována o další dialogy MapTask. Hlavní práce spočívá v digitalizaci a přepisu již pořízených nahrávek a následné značení prozodických událostí.

Dále jsme se zaměřili na digitalizaci zvukových a obrazových dat pořízených v ÚJČ AV. Jedná se o videonahrávky televizních diskusních pořadů, ke kterým v ÚJČ AV vytváří textové přepisy i prozodické značkování. Textové přepisy a značkování jsme převedli do databázové podoby, která umožňuje snadné prohledávání a analýzy všech přepisů. Databáze, dosud založená na dialozích MapTask, tak bude podstatně rozšířena.

T4: Teoretické aspekty komputační lingvistiky, její matematické i lingvistické základy (body A, C, D původního návrhu):

Teoretický výzkum v rámci Centra je neoddělitelně spjat s výše zmíněnými projekty, a to jednak jako předpoklad pro jejich formulaci a teoretický základ pro jejich řešení, jednak tyto projekty přinášejí vedle ověřování platnosti navržených hypotéz i důležité další podněty pro teoretické bádání a pro obohacení daného pojmového rámce. V roce 2003 výzkum pokračoval v následujících bodech:

T4-1 Byly provedeny změny v programech pro trénování pomocí metody maximální entropie tak, aby mohlo být přistoupeno k trénování koeficientů MaxEnt na základě již známého souboru rysů (pravidel). Provedené experimenty ke konci r. 2003 ukázaly, že přes provedené optimalizace je pro češtinu výpočet natolik náročný, že v daném prostoru a s danou technikou ke zlepšení nemůže dojít.

T4-2 V oblasti automatické povrchové syntaktické analýzy, tj. automatická anotace na analytické úrovni PDT pomocí tzv. pravidlového přístupu, bylo vyzkoušeno použití řady pravidlových vzorů pro vývoj pravidlového parseru. Proběhla řada experimentů, které měly za cíl ověřit přínos nejrůznějších modifikací pravidlového přístupu, kde se pravidla určují automaticky.

Byly provedeny první testy, které mají ukázat vhodnost použití tzv. maximálních koster (kořenových stromů) v orientovaných grafech. Úspěšné počáteční experimenty vedly k definici nových a originálních přístupů rozboru vět na analytické rovině.

Paralelně byly zkoumány možnosti kombinace různých již existujících metod. Tento výzkum zatím ukázal, že kombinací tří existujících analyzátorů (Collins, Charniak, Zeman) je možné dosáhnout úspěšnosti nejméně o 1 % vyšší než nejúspěšnější z použitých analyzátorů. Práce v tomto směru pokračují; předpokládáme, že potenciál skrytý v kombinovaných metodách je ještě vyšší.

T4-3 Slovník pro strojový překlad byl obohacen o informace získané automatickým zpracováním lingvistických zdrojů (Euro WordNet, Penn Treebank) a rozsáhlých jednojazyčných dat (American News Texts Corpus, frekvenční slovník češtiny).

T4-4 Pokračovalo se ve studiu hloubkové (tektogramatické) struktury věty na základě dat získaných anotací PDT, a to především těchto aspektů:

- Byly realizovány studie týkající se vztahu povrchových vyjádření valenčních členů (participantů) na analytické rovině k jejich protějškům tektogramatickým.
- Studium valence polysémních sloves se soustředilo na zkoumání vztahu mezi sémantickou charakteristikou slovesa a jeho syntaktickými vlastnostmi; byl vytvořen základ třídění sloves podle jejich syntakticko-sémantických charakteristik.
- Valenční pozice sloves a substantiv byly zkoumány z hlediska asymetrií mezi tektogramatickou (hloubkovou) rovinou a povrchovou strukturou; rozlišují se nulové povrchové podoby odpovídající různým typům valenčních doplnění: nula odpovídající
  (a) všeobecnému aktantu (s lemmatem Gen v TGTS; *V novinách se píše* (Gen.ACT)),
  (b) „nespecifikovanému" aktantu (s lemmatem Unsp v TGTS; *V novinách píší* (Unsp.ACT)),
  (c) deikticko-anaforické elementy plynoucí z kontextu nebo situace u sémanticky obligatorních směrových určení (s lemmatem *sem*, *tam* apod. a příslušným funktorem; *Moji přátelé právě odjeli* (tady.DIR1; odpovídá určení *odtud*)).
- Studium konkurence shodného a neshodného substantivního atributu se soustředilo zejména na problematiku atributu vyjádřeného substantivem; v této souvislosti byla studována otázka shody a neshody dvou substantiv ve větě vůbec, přičemž tato „substantivní shoda" (např. *ministr Dostál*) byla odlišena od tradičně chápané shody adjektivní (atributivní).
- V oblasti aktuálního členění věty bylo dopracováno zejména: přesnější vymezení pojmů základ, kontrastivní téma, ohnisko, vlastní ohnisko, kvaziohnisko, kontextová zapojenost; zachycení výpovědní dynamičnosti v substantivní skupině a v kontextově zapojené části výpovědi; zachycení aktuálního členění struktur, které jsou na analytické rovině neprojektivní; zachycení aktuálního členění nevětných struktur; zachycení aktuálního členění otázky; studium kontrastu, a to z hlediska vztahu k sémantické stavbě věty a strukturaci textu s cílem jeho podrobnější klasifikace, a dále z hlediska role kontrastu při vzniku příznakových slovosledných variant.
- V oblasti diskurzu se zkoumají možnosti automatického zachycení struktury diskurzu na základě dříve formulovaného předběžného algoritmu přiřazování stupňů aktivovanosti jednotlivým členům věty. Byla zpřesněna heuristická pravidla pro přiřazování stupňů aktivovanosti jednotlivým členům vět a byla zahájena formulace algoritmu a jeho počítačové implementace založená na anotování Pražského závislostního korpusu.

Na spoluřešitelském pracovišti (UJČ) byly na základě dat z PDT zkoumány následující problémy:

- vývojová dynamičnost v nominální skupině: jak se mění
  a) poměr (přibývající) prepozice a (ubývající) postpozice adjektivních atributů;
  b) poměr prepozice a postpozice se zřetelem na rozměr rozvíjejících členů;
  c) konkurence posesivního adjektiva a genitivu substantiva;
  d) užívání neslovních / slovních nesklonných výrazů v prepozici (typ *fotbalová Gambrinus liga*);
  e) vzájemná poloha genitivního a předložkového atributu k témuž řídícímu jménu (typ *překlad do ruštiny kolektivní monografie*);
- slovosled věty a aktuální členění větné: poloha verba finita vůbec, poloha rematického verba finita, poloha verba finita ve vedlejších větách; statistika začátku a konce věty: zakotvenost věty v předcházejícím větném kontextu.

T4-5 Byly prostudovány různé přístupy k logické sémantice z hlediska případné formulace logicko-

sémantické reprezentace věty jako další stupeň stratifikačního scénáře anotace Pražského závislostního korpusu. Byly rovněž prostudovány různé přístupy k lexikálně sémantické analýze a různé systémy kategorizace lexikálních významů (tzv. ontologie). K této problematice byl uspořádán mezinárodní seminář za účasti osmi předních světových expertů v této oblasti (viz bod T6-3). Déle probíhaly neformální konzultace se zahraničními odborníky (zejména s Universitou v Saarbruecken) týkající se možného zachycení významu často se vyskytujících, avšak dosud neprozkoumaných typů propozic.

T4-6 V rámci studia alternativních přístupů jsme se nově zaměřili i na studium využití nelineárních systémů pro zpracování přirozeného jazyka. Po publikaci několika experimentů z tohoto v komputační lingvistice zcela nového přístupu jsme se soustředili na budování sítě budoucí spolupráce v této oblasti a tím na možné pokračování bádání v této interdisciplinární sféře (viz bod 4. Spolupráce Centra).

T5: Vyvíjení některých menších aplikačních systémů (bod F).

T5-1 V rámci projektu česko-slovenského překladu Česílko byly texty z PDT experimentálně přeloženy za použití nejnovější verze slovníků a českého taggeru. Pracovalo se na dalším rozšiřování překladových slovníků.

Dále byl vyvinut experimentální systém strojového překladu z češtiny do litevštiny. Architektura systému je rozšířením původní implementace systému strojového překladu z češtiny do slovenštiny a do polštiny (viz průběžné zprávy CKL za léta 2001 a 2002), komponentu částečného syntaktického transferu. Fáze analýzy sestává z morfologické analýzy, desambiguace a parciální syntaktické analýzy. Transfer spočívá v úpravě nejčastějších jevů odlišných v obou jazycích, např. slovosledu jmenných skupin a morfologie sloves. Fáze syntézy se skládá se z linearizace syntaktických struktur a morfologického zpracování výstupu. Dosažené výsledky prokazují výrazné zlepšení kvality překladu po přidání komponenty částečné syntaktické analýzy a transferu. Dosažená přesnost překladu (vážený průměr) měřená systémem Trados Translator's Workbench je 87,6% (pro srovnání, pro polštinu bez transferu je přesnost 71,4%).

T5-3 V rámci projektu "Sémantické modely textů" byla připravena testovací data. Byla implementována většina dílčích modulů (tj. předzpracování vstupních dat, extrakce kolokací, analýza shluků, skládání virtuálních konceptů) a pracuje se na jejich integraci. Byly provedeny základní experimenty. Pro dokončení projektu je nyní potřeba natrénovat parametry modelu pomocí nepříliš rozsáhlých anotovaných dat, na jejichž ruční anotaci pomocí speciálně připraveného softwaru se již začalo pracovat. Poslední fází bude provedeni evaluace modelu, rovněž s využitím anotovaných dat.

T6: Přípravy velkých mezinárodních akcí

T6-1 Jarní škola Viléma Mathesia. Ve dnech 9.- 22. března proběhl osmnáctý cyklus mezinárodní jarní školy Viléma Mathesia (VMC), na kterém přednášelo devět zahraničních profesorů a účastnilo se 55 zahraničních i českých studentů. Poté byla vypracována zpráva pro grantové agentury a uzavřeno vyúčtování za tento ročník. Dále byla zahájena příprava programu devatenáctého cyklu kurzů, který se bude konat na jaře roku 2004.

T6-2 Mezinárodní kongres lingvistů. Ve dnech 24.-29.7.2003. se v Praze pod patronátem mezinárodní organizace Comité International Permanent des Linguistes (CIPL) uskutečnil Mezinárodní kongres lingvistů (CILXVII), jehož bylo Centrum hlavním spolupořadatelem (zajišťovalo průběh kongresu, a to jak po stránce organizační, tak i po stránce programové a společenské). Kongresu se zúčastnilo 436 účastníků, z toho 303 regulérních, 105 studentů a 26 neplatících účastníků (členové CIPL, zvaní).

Kongres se setkal s velice kladným ohlasem jak po stránce odborné a organizační, tak i po stránce společenské, o čemž svědčí reakce od účastníků a členů exekutivy CIPL. Kongres byl, co se jeho úrovně týká, hodnocen jako jeden z nejlepších od počátků jejich konání.

Organizátoři se, bohužel, potýkali se značnými finančními problémy, které byly způsobeny trojnásobně nižší účastí než bylo CIPLem oznámeno. Díky úsilí lokálního organizačního výboru při zajišťování sponzorů je tento negativní finanční dopad téměř vyrovnán.

K CILXVI byl vydán sborník abstraktů (tištěná forma) a sborník příspěvků (CD-ROM):
- XVII International Congress of Linguists. Abstracts. (Ed. Hajičová, E.), CKL, 2003.
- Proceedings of the XVII International Congress of Linguists. (Eds. Mírovský, J., Kotěšovcová, A., Hajičová, E.), CKL, 2003.

T6-3 Mezinárodní seminář o klasifikaci lexikálních významů (tzv. ontologii). Seminář proběhl ve dnech 8. a 9. prosince a k přednáškám bylo pozváno 8 zahraničních profesorů – významných odborníků v této oblasti (F. Jelinek a J. Pustejovsky z USA, S. Atkins, L. Guthrie, Y. Wilks a P. Hanks z Velké Británie, N. Calzolari a G. Basili z Itálie). Semináře se zúčastnili jak pracovníci CKL, tak i další pracovníci MFF UK, FF UK, UJC AV, Masarykovy univerzity v Brně, pracovníci Jazykovedného ústavu SAV v Bratislavě a další. Vedle pozvaných přednášek byl velký prostor věnován diskusním blokům k jednotlivým tématům a otázkám. Závěry semináře se stanou východiskem pro koncipování jednoho z dalších projektů v rámci Pražského závislostního korpusu, který by měl být formulován pro období po roce 2004.

Další aktivity Centra v roce 2003

Nové projekty řešené v rámci Centra komputační lingvistiky

- **Prague Arabic Dependency Treebank**
  Centrum komputační lingvistiky se zapojilo do zatím neformální spolupráce s Ústavem formální a aplikované lingvistiky MFF UK a Ústavem srovnávací jazykovědy FF UK, jejímž cílem je budovámí Pražského závislostního korpusu pro arabštinu (Prague Arabic Dependency Treebank). Tento projekt se začal připravovat již v letech 2001/2002, v roce 2003 došlo k výraznému zintenzivnění spolupráce a k zapojení dalších pracovníků Centra. Zdrojový korpus dat a morfologický analyzátor poskytlo Linguistic Data Consorcium (LDC), University of Pennsylvania.
  V roce 2003 se pracovalo na morfologickém anotování pomocí nástroje výhodně získaného od LDC (anotováno 60 000 slov). Dále se připravovaly podklady pro analytické značkování.
  Dále byl v Praze organizován seminář s pracovníky Linguistic Data Consortium, University of Pennsylvannia, PA, USA (viz též dále, semináře a workshopy), se kterými CKL na projektu úzce spolupracuje, a všech zúčastněných českých institucí, kde byla stanovena strategie dalšího postupu a koncepce dlouhodobé spolupráce. Bylo dohodnuto, že se bude používat jednotný systém morfologické analýzy, který se v Praze pouze upravuje, a jednotný nástroj vyvinutý v LDC na manuální anotaci. Syntaktická anotace bude prováděna a řízena oběma pracovišti nezávisle, avšak bude se pracovat na konverzi mezi oběma použitými systémy tak, aby se objem anotovaných dat mohl v obou systémech zdvojnásobit. Cílem je anotovat text o délce 500 tisíc slov, z toho část ve formě tzv. paralelního koprusu (arabština/angličtina). Pro naše instituce tento projekt představuje nově otevřené možnosti spolupráce na jazykových projektech s podobnou tématikou řešených na pracovištích v USA.
- **Zpracování psaného kulturního dědictví**
  Ve spolupráci se studenty MFF UK a institucemi zabývající se zpracováním staroslověnských a církevněslovanských textů (viz níže, bod 4. Spolupráce Centra) vzniká v CKL komplexní systém pro lingvisticky orientované počítačové zpracování textů staroslověnských a církevněslovanských památek.
  Byla vytvořena sada nástrojů pod názvem ACT (Annotated Corpora of Text), což je jazykově nezávislý systém skládající se z následujících modulů:
  a) ACT Client a ACT Server vytváří základ klient-server architektury s napojení na databázi zpracovaných rukopisů (cca. 600 000 lematizovaných a slovnědruhově označkovaných forem, vč. další anotace)
  b) ACT Client Light (modul pro mimosíťovou práci s dokumenty)
  c) ACT Distiller (automatizovaný import existujících ručně excerpovaných lexikografických kartiček)
  b) ACT Web (zpřístupňuje data přes webové rozhraní).
  CKL tímto přispívá k budování informačních technologií pro zpracování památek psaného kulturního (slovanského tedy i českého) dědictví.

Přednáškové pobyty předních zahraničních profesorů.

- Na pozvání CKL připravil prof. E. Charniak, Brown University, USA pro studenty a pracovníky CKL intenzivní a velmi přínosný **kurs "Statistical parsing of natural language"**.
- Prof. Greenberg (USA) přednesl přednášku o automatickém rozpoznání prozodie.
- Prof. M. Palmer seznámila studenty a pracovníky CKL s anotací anglického korpusu PropBank, který lze chápat jako analogii tektogramatického značkování PDT pro angličtinu. V PropBanku je nyní zachycována argumentová struktura sloves.
- Dále Centrum zabezpečilo příjezd deseti předních odborníků v oblasti komputační lingvistiky na **Světový kongres lingvistů** (např. prof. W. Dressler, Rakousko, prof. F. Kiefer, Maďarsko, prof. Ch. Lehmann, Německo, prof. P. Ramat, Itálie, prof. J. Perrot, Francie), viz bod T6-2.
- CKL zabezpečilo též příjezd šesti významných odborníků v oblasti lexikální sémantiky, kteří se podíleli na **semináři o klasifikaci lexikálních významů, tzv. ontologií** (např. prof. L. Guthri, Velká Británie, prof. Pustejovsky, USA, prof. P. Hanks, SRN) viz bod T6-3.

Výjezdní semináře:

- **Nové Hutě na Šumavě**, 12.-17.1.2003
  Mladí pracovníci Centra přednesli referáty jednak jako zprávy o výsledcích dosažených v rámci jejich doktorského studia, jednak jako zprávy o úkolech plněných v souvislosti s projekty Centra (např. M. Holub, *Konceptově orientované vyhledávání informací a sémantické modely textu*; M. Razímová, *Německé tektogramatické stromy*; L. Kučová, *Zachycení koreference v PDT a problémy s tím spojené*; V. Řezníčková, *Lemata doplněných uzlů v závislosti na typu koreferenčního vztahu,* L. Uhlířová, *Gramatika v korpusu, korpus v gramatice*). Jednacím jazykem byla zčásti angličtina, což zejména pro mladé pracovníky představuje vítaný trénink v anglické prezentaci vlastních výsledků (např. M. Lopatková, *Valency in PDT and Valency Lexicon*; A. Böhmová, V. Honetschlager, *Semi-automatic tectogrammatical tagging*; O Smrž, *Prague Arabic Treebanking*). Velikým přínosem byla též aktivní účast prof. E. Charniaka z USA (*Parsing Czech, Parsing English*) a jeho doktorandky Heidi Fox (*Using English Parser for French-English Alignment*).
- **Doubice u Děčína**, 26.-29.09
  Společného pobytu bylo využito k hodnocení prvního půlroku práce na projektech a k diskusi o výhledech na rok 2004 a směrech dalšího pokračování projektu. Po půlročním provozu byla zhodnocena funkčnost nových www stránek CKL a možnosti jejich dalšího zlepšování (databáze publikací, odpovědnosti jednotlivých pracovníků).

Uskutečněné a plánované workshopy:

- **Prague-Penn Arabic Treebanking Workshop**, 21.-28.5. 2003
  CKL uspořádalo pracovní workshop se sedmi kolegy z Linguistic Data Consortium, University of Pennsylvania, kteří tvoří vedení tzv. Arabic Treebanking Group, spolupracující úzce s českou stranou. Na workshopu byly představeny výsledky české části projektu, navržena nová řešení a dohodnuty kroky dalšího vývoje.
  Workshop byl vysoce hodnocen z hlediska věcného i organizačního. Za českou stranu byly představeny úspěchy projektu (referát na EACL'03, postup anotací, softwarové nástroje Netgraph a TrEd), navržena nová řešení (reimplementace analyzátoru AraMorph, Prague Markup Language, atd.) a dohodnuty kroky dalšího vývoje (sdílení dat i anotačních nástrojů, křížové anotace, systematické porovnání metodologií).
- **Návrh workshopu „Arabic Treebanking Workshop" na LREC 2004**
  Pracovníci CKL připravili návrh na pořádání zvláštního workshopu při konferenci LREC 2004. Organizační výbor předpokládá, že součástí workshopu bude zvaná přednáška věnovaná projektu Prague Arabic Dependency Treebank a jeho využití pro úlohy NLP. Znění návrhu je k dispozici na http://ckl.mff.cuni.cz/smrz/LREC04/Arabic_Workshop_Proposal.htm. Tento návrh na workshop byl přijat, workshop proběhne 24. května 2004.

Přínos zahraničních cest

Přehled zahraničních cest pracovníků CKL (níže) i oddělený soupis zahraničních cest mladých

pracovníků a doktorandů (viz bod 6.) jasně ukazuje, že o výsledky výzkumu CKL je v zahraničí velký zájem: většina cest totiž pokrývá buď účast na mezinárodní konferenci s pozvaným nebo přijatým referátem, nebo aktivní účast na mezinárodním vysoce ceněném workshopu, nebo přednáškový pobyt na některé zahraniční univerzitě. Řada pobytů byla částečně hrazena zahraničním grantem nebo přímo zvoucí organizací, takže z prostředků Centra byl hrazen jen nutný doplněk výdajů.
Během zahraničních cest bylo uzavřeny nové kontakty a byla připravována náplň vědeckých projektů v rámci Evropské Unie, ke kterým bylo CKL přizváno, více viz bod 4. níže.

Přehled zahraničních cest uskutečněných z prostředků CKL
V následujícím soupisu uvádíme zahraniční cesty pracovníků CKL, cesty mladých pracovníků do 35 let jsou uváděny zvlášť (viz bod 6).

Eva Hajičová

- **Londýn, Lancaster, Anglie, březen 2003**
  Účast na mezinárodní konferenci Corpus Linguistics 2003 v Lancasteru; předsedání jedné ze sekcí konference.
- **Budapešť, Maďarsko, duben 2003**
  Účast na mezinárodní konferenci EACL 2003.
- **Paříž, Francie, květen 2003**
  Příprava projektu Barrande na léto 2004-2005; příprava náplně mezinárodního projekdu MULTIP v rámci programu EU, 6. Framework.
- **Ženeva, Švýcarsko, květen 2003**
  Jako členka programového výboru mezinárodní konference EUROSPEECH účast na zasedání výboru; jako členka exekutivy Int. Speech and Communication Association účast na zasedání rady.
- **Göteborg, Švédsko, červen 2003**
  Přednáškový cyklus na letní škole slovanských studiií, pořádané univerzitou v Göteborgu
- **Sapporo, Japonsko, červenec 2003**
  Účast na mezinárodní konferenci o počítačové lingvistice pořádané světovou org. ACL; účast na 2 workshopech po konferenci (o anotaci a o výzkumu nad velkými jazykovými daty). Předsedání jedné ze sekcí mez. konf. ACL.
- **Lublaň, Slovinsko, srpen 2003**
  Účast na 13. Mezinárodním kongresu slavistů. Přednesení referátu.
- **Terst, Itálie, srpen 2003**
  Přednáškový cyklus na 9. Evropské letní škole o formální gramatice.
- **Paříž, Francie, srpen 2003**
  Účast na konferenci - projekt ENABLER.
- **Ženeva, Švýcarsko, srpen 2003**
  Účast na konferenci Eurospeech 2003, předsedání sekce, (jako členka progr. výboru). Účast na zasedání rady ISCA (jako členka výkonného výboru).
- **Bratislava, Slovensko, září 2003**
  Účast na slavnostním zasedání k 60. Výročí trvání Ústavu (pozvána jako předsedkyně Pražského lingvistického kroužku).
- **Paříž, Francie, listopad 2003**
  Účast na konferenci LangTech 2003 a na organizační schůzi navrhovatelů projektu ERA-Net in Language and Speech Technology v rámci 6. programu EU.
- **Lipsko, Německo, listopad 2003**
  Účast na mezinárodní konferenci o formálním popisu slovanských jazyků, přednesení referátu.
- **Pisa, Itálie, listopad 2003**
  Účast na zasedání Vědecké rady Centra komputační lingvistiky (ILC) v Pise (jako předsedkyně rady).

Jaroslava Hlaváčová

- **Brighton, Velká Británie, prosinec 2003**
  Účast na workshopu 2nd International Workshop on Dictionary Writing Systems.

Pavel Ircing

- **Tokio, Japonsko, duben 2003**
  Na workshopu Workshopu SSPR 2003 při konferenci ISCA&IEEE prezentoval referát *Recognition of Spontaneously Pronounced TV Ice-Hockey Commentary.*
- **Ženeva, Švýcarsko, září 2003**
  Na konferenci Eurospeech 2003 prezentoval referát Fitting *Class-Based Language Models into Weighted Finite-State Transducer Framework.*

Vladislav Kuboň

- **New Orleans, Spojené státy, září 2003**
  Účast na konferenci Machine Translation, referát *A Simple Multilingual Machine Translation Systém.*
- **Londýn, Velká Británie, září 2003**
  Účast na konferenci Translation and computers organizované ASLIB.
- **Paříž, Francie, listopad 2003**
  Účast na konferenci LangTech 2003.

Markéta Lopatková

- **Vysoké Tatry, Slovensko, září 2003**
  Účast na konferenci ITAT 2003, přednesení referátu *Testování konzistence a úplnosti valenčního slovníku českých sloves* (spoluautor Z. Žabokrtský).
- **Bratislava, Slovensko, říjen 2003**
  Přednáška v Ústavu slovenského národného korpusu, JULŠ, SAV *Valence a Pražský závislostní korpus (PDT)* (společně s J. Panevovou).

Josef Psutka

- **Orlando, USA, červenec 2003**
  Na konferenci SCI'2003 přednesl referát *Automatic Transcription of TV Ice-Hockey Commentary.*
- **Ženeva, Švýcarsko, září 2003**
  Na konferenci Eurospeech 2003 prezentoval referát na téma *Large Vocabulary ASR for Spontaneous Czech in the MALACH Project*.

Petr Sgall

- **Londýn, Lancaster, Anglie, březen 2003**
  Účast na mezinárodní konferenci Corpus Linguistics 2003 v Lancasteru.
- **Terst, Itálie, srpen 2003**
  Letní škola o závislostní gramatice – cyklus přednášek.

František Štícha

- **Mannheim, Německo, 1. – 19. 6. 2003**
  Studijní pobyt v Ústavu pro německý jazyk v Mannheimu, jehož účelem bylo seznámit se s novou verzí vyhledávacího programu COSMAS II v jeho řádkové i grafické podobě a s možnostmi získávání informací o frekvenci a textové distribuci jazykových struktur ve velkých korpusech přesahujících jednu miliardu slovních forem.

Ludmila Uhlířová

- **Trier, Německo, říjen 2003**
  Účast na konferenci 4th Colloqium of Quantitative Linguistics, přednesení referátu *Zipf´s Law for Pair of Words.*
- **Lyon, Francie, září 2003**
  Účast na konferenci SLE Lyon, 4.-7.9.2003, přednesení referátu *´This/That´in Discourse Structure: An evidence from the Czech National Corpus.*

Zdeňka Urešová

- **Bratislava, Slovensko, září 2003**
  Jednání v JÚLŠ SAV SR o MVTS a demonstrace vyhledávacích programů.
- **Prešov, Slovensko, prosinec 2003**
  Seznamování budoucích anotátorů Slovenského národního korpusu s anotačními nástroji vyvinutými v CKL, prezentace syntaktického značkování PDT na analytické rovině, instalace grafového editoru TrEd a školení studentů v práci s tímto anotačním nástrojem.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1,110 00, IČO: 00216208

Změny v projektu

a) změny v projektu

V těsné návaznosti na cíle projektu došlo v roce 2003 k následujícímu rozšíření činnosti CKL:
Kromě několika menších projektů souvisejících s dosavadní činností Centra (ČAK, bod T1-9, využití nelineárních systémů, bod T4-6) se intenzivně rozvíjela práce na dvou nových mezioborových projektech zaměřených na získávání jazykových dat a vyvíjení souvisejících nástrojů. Tyto projekty souvisí s vědeckým zájmem mladých pracovníků CKL, který je 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).

- **Prague Arabic Treebank** (viz výše, Další aktivity Centra v roce 2003)
- **Zpracování psaného kulturního dědictví** (viz výše, Další aktivity Centra v roce 2003)

b) přehled nesplněných úkolů

V náplni vědecké práce nedošlo k žádným změnám ani přesunům; pokud jde o některé menší přesuny finančních prostředků, viz o tom bod *Podrobná specifikace a zdůvodnění jednotlivých položek finančních prostředků projektu* níže.

Výstupy CKL – vyvinutá data a související nástroje

Konkrétních výsledky a výstupy získaných v rámci realizace projektu jsou uvedeny v Příloze 1 **Seznam dat a souvisejících nástrojů získaných v rámci realizace projektu.**

Publikační činnost pracovníků CKL

Seznam publikací, které vznikly za podpory CKL a byly vydány v r. 2003, a prací připravených k publikaci je uveden v Příloze 2 **Publikace**.

2. Personální a organizační zabezpečení činnosti Centra

V roce 2003 nedošlo k významnějším změnám v personálním a organizačním zabezpečení Centra. Na řešení projektu se podílí 40 pracovníků, kteří dohromady zastávají 24,7 plných úvazků (rok 2000 - 28 prac.; r. 2001 - 42 prac.; r. 2002 - 41), a dále 21 studentů zastávajících 10,75 plných úvazku.

Složení pracovního týmu je z hlediska kvalifikace ve vztahu k náplni v Centru vyvážené. Pracovní tým CKL se skládá ze čtyř profesorů, devíti vědeckých pracovníků a 24 odborných pracovníků. Jeho činnost zajišťují tři techničtí pracovníci.

S CKL dlouhodobě spolupracuje řada studentů magisterského studia (20-24 studentů, podle časových možností) a další odborní pracovníci (3).

Věkovým zastoupením svých pracovníků CKL splňuje podmínku zaměstnávat především mladé vědecké a odborné pracovníky – 25 pracovníků CKL je mladších než 35 let, dohromady zastávají 17,1 plných úvazků (tj. více než 69%), další objem práce odvádějí studenti magisterského studia (všichni do 35 let).

Kvalifikace, prac. zař.	počet	prům. věk	vážený věk	úvazek
profesoři	4	65	68	2,6
docent	1	41	41	0,2
vědečtí pracovníci	8	49	37,4	2,5
odborní pracovníci	24	30	30	16,6
technická podpora	3	40	39	2,8
Celkem	**40**			**24,7**
Studenti	**21**			**10,75**

Celková výše úvazků v CKL zůstala na stejné úrovni jako v roce 2003. Z důvodu personálního zabezpečení organizace Mezinárodního kongresu lingvistů (CKL bylo jeho hlavním spolupořadatelem, viz bod T6-2) byli někteří dlouhodobě spolupracující studenti zaměstnání na větší část roku na částečné úvazky. Z ostatních osobních nákladů byla hrazena krátkodobá spolupráce studentů zajišťujících organizaci kongresu.

Pracovník CKL Jan Hajič v roce 2003 po úspěšném habilitačním řízení získal titul docent.
Pracovník CKL František Štícha v říjnu 2003 úspěšně prošel habilitačním řízením na vědecké radě FF UK.
Pracovník CKL Pavel Ircing též v roce 2003 dokončil a odevzdal disertační práci (viz seznam publikací), která bezprostředně souvisí s problematikou řešenou v rámci Centra. Práce je v současné době v oponentním řízení.
Dále pět doktorandů zaměstnaných v CKL úspěšně složilo doktorské zkoušky, které jsou nezbytným předpokladem pro podání disertační práce.
Vedoucí CKL Eva Hajičová dostala v roce 2003 Medaili I. stupně ministryně školství, mládeže a tělovýchovy za vynikající vědecké a pedagogické výsledky v oblasti komputační lingvistiky.

3. Přístrojové vybavení a technické zabezpečení činnosti Centra

MFF UK

Investiční prostředky pro rok 2003 byly přesunuty a proinvestovány v předcházejícím roce, takže nebyly provedeny žádné nákupy výpočetní techniky ani upgrady. Jak jsme předpokládali, nebylo nutné provádět žádné větší opravy a i ty menší byly zpravidla pokryty zárukou.
Během roku 2003 sice technologický vývoj pokračoval obvyklým tempem, nicméně dá se říci, že díky vhodně naplánovaným nákupům byla výpočetní síla Centra po celou dobu zcela dostatečná. Jelikož došlo ke zpoždění ve vývoji 64-bitových procesorů Opteron, získali jsme nákupem stroje s procesory Itanium2 roční náskok v přechodu aplikací na 64-bitovou platformu.
V následujícím roce tak budeme moci lépe využít strojů, postavených právě na této nové řadě procesorů, jak jsme dlouhodobě plánovali.

Lze říci, že vybavení Centra je využíváno v maximální míře a vyhovuje současným nárokům. Posílení výpočetní a datové základny předpokládáme začátkem léta. V tomto období také započne částečná obnova nejstaršího vybavení.

ZČU

V roce 2003 byly na účet spoluřešitele (ZČU) převedeny hlavním řešitelem (MFF UK) investiční prostředky dle původního plánu ve výši 300 tis. Kč. Za zmíněné investiční prostředky byla provedena inovace výpočetních stanic CKL:
Dvě velmi výkonné pracovní stanice Fujitsu-Siemens Celsius R610: Systemboard D1357, 2x Intel Xeon 2,40 GHz /533MHz / 512kB, 2 GB DDR SDRAM, Pevný disk 40 + 120 GB, Grafika Celsius Quadro4 550, Floppy mechanika 3.5" / 1.44MB, Mechanika DVD-ROM 16x/40x, Mechanika CD-RW 48x/16x/48x, Klávesnice, Myš wheel, plochý LCD Display FSC P-19 19", rozšířená záruka 3 roky, 48h on-site, SW OEM Win XP Pro.
Celková cena 300 tis. Kč

Pozn.:
Doplnění PC pracovních stanic bylo realizováno z důvodů potřeby permanentního a časově velmi náročného trénování akustických a jazykových modelů. Vzhledem k tomu, že pracovníci Centra komputační lingvistiky participují na řešení rozsáhlého projektu „MALACH“, kde jsou řešeny úlohy akustického a jazykového modelování, a to vedle češtiny i několika dalších evropských jazyků, je pro tyto účely zapotřebí značný výpočetní výkon. Navíc, na řešení úloh Centra se podílí stále větší počet studentů magisterského a doktorského studia, kteří plně využívají sice starší, ale stále plně funkční investiční prostředky zakoupené v minulém období.

ÚJČ

Pro nové pracovníky jazykové poradny, pro potřeby spolupráce ÚJČ s externími pracovníky v rámci spolupráce ÚJČ a CKL a pro potřeby ukládání dílčích korpusů a práce s nimi mimo pracoviště byly zakoupeny dva notebooky s jednou cestovní přenosnou tiskárnou v celkové hodnotě 150 000 Kč:
2 x notebook ACER TravelMate + příslušenství (brašna, myš)
tiskárna HP DJ 450cbi
Celková cena 150 tis. Kč

4. Spolupráce Centra

a) Rozvoj odborné spolupráce v rámci ČR, s ostatními zakládajícími a spolupracujícími organizacemi ve sledovaném období

Dále se prohloubila spolupráce s Fakultou aplikovaných věd ZČU v Plzni (návrh projektu Spoken Translation: Merging Rules and Statistics, viz níž bod b), s Laboratoří přirozeného jazyka na FI MU Brně (problematika korpusů, porovnání přístupů k zachycení valence sloves), s Ústavem Českého národního korpusu (nadstandardní přístup k datům ČNK) a s Ústavem teoretické a komputační lingvistiky (propojení statistických a pravidlových přístupů k morfologické disambiguaci, výpočtová složitost).

V rámci vytváření Pražského arabského závislostního korpusu (viz výše, Další aktivity Centra v roce 2003) probíhá úzká (neformální) spolupráce s Ústavem srovnávací jazykovědy FF UK.

V novém projektu *Zpracování psaného kulturního dědictví* spolupracuje CKL se Slovanským ústavem AV ČR.

Další intenzívní odborná spolupráce byla navázána především s řešiteli projektu GAČR 405/03/0377, *Možnosti a meze gramatiky češtiny ve světle Českého národního korpusu*.
CKL dále spolupracuje s řešiteli nových projektů GAČR 405/03/0913, *Very Large Language Corpora and their Automatical Analysis* a GAČR 405/03/0914, *Machine Translation of Economic Texts from Czech to English*.

b) Nová zapojení do mezinárodních struktur ve sledovaném období

Slovensko
V rámci projektu česko-slovenského překladu byla navázána formálně-smluvní spolupráce s Jazykovedným ústavem LŠ v Bratislavě v oblasti vývoje a údržby slovenského morfologického slovníku a s partnerským pracovištěm na Pedagogické fakultě Komenského v Bratislavě (viz T5-1).
Byly získány dodatečné prostředky (mobilita pracovníků) v rámci programu MŠMT KONTAKT na spolupráci se slovenským partnerem, a to na r. 2004-2005, s cílem zintenzívnit přenos poznatků získaných v rámci CKL na obdobný projekt nyní řešený v JÚLŠ v Bratislavě a na FF v Prešově.

Evropské projekty
V průběhu roku bylo rovněž připravována náplň několika projektů programu EU, 6. Framework:

- **Multilingual Textual Information Processing (MULTIPLE)**
  Network of Excellence, organizuje Centre de recherche en linguistique et traitement automatique des langues Lucien Tesnière, Université de Franche-Comté, Francie.
- **European Lexical Infrastructure and Technology (ELITE)**
  Network of Excellence, organizuje prof. N. Calzolari, ILC, Pisa, Itálie.
- **Semantic multilingual architecture for information extraction services (SMARTIES)**
  Typ projektu STREP, vedoucí organizace projektu ComNet Media, německá firma specializující se na vyhledávání elektronických informací.
- **Mobile Confidants and Companions (MC2)**
  Typ projektu IP, vedoucí organizace projektu University of Sheffield, GB, oddělení Lingvistiky, ved. Prof. Yorick Wilks.

V současné době jsou podány tyto návrhy na industriální a mezinárodní spolupráci:

- **Multimodální Aplikační Rozhraní, Technologie a Aplikace (MARTA)**
  Projekt programu Tandem, Ministerstvo průmyslu a obchodu ČR, vedoucí organizace projektu ISC Communication, a.s., kromě CKL MFF UK spolupracuje IBM ČR, ČVUT Praha, ÚTIA AV ČR.

- **European platform for localisation quality** (**ELOQUATE)**
  Projekt 6.FP EU, typ projektu CRAFT (podpora SME), **v**edoucí organizace projektu: Saarland University, Německo, prof. Hans Uszkoreit. Dále participují: Text & Form GmbH, Skřivánek (ČR), PuntoLOC (Šp.), Comunicación Multilingüe (Šp.), Acrolinx GmbH (celkem 4 SME a 3 univerzity).
- **Spoken Translation: Merging Rules and Statistics**
  Projekt programu "New Innovation Team Eastern Europe", vedoucí pracoviště projektu CKL (UK Praha a ZČU Plzeň). Projekt spolupráce s Výzkumným centrem firmy Samsung**.**
- **ERA-Net in Language and Speech Technology, LangNet**
  "Coordination of research activity" v rámci programu "Integrating and Strengthening the European Research Area", ved. dr. J. Mariani.

Pokračovali jsme v dosavadní mezinárodní spolupráci, jak jsme o ní podávali zprávy v minulých letech (např. ENABLER).

V rámci nového projektu *Zpracování psaného kulturního dědictví* (viz výše, Další aktivity Centra v roce 2003) CKL navázalo intenzivní neformální spolupráci zejména s Bulharskou akademií věd v Sofii, s Ústavem makedonského jazyka ve Skopje v Makedonii, s výzkumným centrem Istituto di Lingustica Computazionale v Itálii, a s Univerzitou v Pise v Itálii.
Byl vypracován a přijat mezinárodní projekt „Transfer of Knowledge" mezi CKL a Bulharskou akademií věd, kde CKL má funkci školitele. V současné době probíhá sestavování návrhu výměny (v rámci projektu ERASMUS) mezi Fakultou informatiky Univerzity v Pise (Itálie) a CKL, jejíž primárním cílem je rozšíření získaných zkušeností pracovníků CKL i za hranice České republiky a podpoření mezinárodní výměny znalostí a zkušeností.
Pracovník Centra RNDr. Kiril Ribarov se stal spolupředsedou Komise pro počítačové zpracování středověkých rukopisů a starých knih, při Mezinárodním komitétu slavistů.

USA
V roce 2003 došlo k dalšímu prohloubení spolupráce s Linguistic Data Consortium, University of Pennsylvania, které bylo zaměřeno zejména na výměnu nástrojů pro anotaci, viz výše Další aktivity Centra v roce 2003 a níže bod d)).

V rámci prací na projektu MALACH (viz bod T3-1) CKL dále intenzivně spolupracuje s Johns Hopkins University v Baltimore, University of Maryland, IBM (Human Language Technologies Group, Yorktown) a Visual History Foundation v Hollywoodu.

V rámcu projektu zkoumajícího možností využití nelineárních systému pro zpracování přirozeného jazyka byla navázána zatím neformální spolupráce s University of California San Diego (UCSD).

c) Rozvoj spolupráce s aplikační sférou a v rámci regionu

V roce 2003 byly předány k užívání nástroje, data a postupy vyvinuté v CKL.

Na pracovišti Slovenského národního korpusu (SNK) v Bratislavě byl předveden vyhledávací program NetGraph (program sloužící k vyhledávání struktury vět v PDT). Dále proběhla diskuse o projektu slovenské morfologie, týkající se zejména morfologické anotace SNK, a pracovníkům SNK byl předán slovenský morfologický slovník.

Na FF PU v Prešově proběhlo školení budoucích anotátorů Slovenského národního korpusu. Školení spočívalo v popisu a ukázkách syntaktických anotací PDT, a to jak anotací na analytické rovině, tak anotací na tektogramatické rovině. Pracovníkům FF UP byly rovněž předány do zkušebního provozu anotační nástroje vyvinuté v CKL.

Významným důsledkem pražského workshopu Prague-Penn Arabic Treebanking Workshop (viz výše, Další aktivity Centra v roce 2003) byla pracovní cesta Romana Ondrušky a Jiřího

Mírovského, autorů systému Netgraph pro efektivní vyhledávání v grafech a strukturách nad jazykovými daty, do Linguistic Data Consortium ve dnech 13.10. až 27.10. 2003. Programem jejich práce byla instalace systému na tamní síť, převod korpusů do formátu vhodného pro vyhledávání i provedení některých úprav a vylepšení podle požadavků LDC. Autoři též přednesli souhrnnou prezentaci pro celé pracoviště a provedli podrobné školení.

Pro vědecké účely byla uvolněna první verze komplexně anotovaného valenčního slovníku českých sloves VALLEX 1.0, který je vyvíjen v souvislosti s anotací na tektogramatické rovině PDT. Vzhledem ke kladným ohlasům předpokládáme jeho využití v aplikacích souvisejících s automatickým zpracováním češtiny.

d) Způsob využívání výsledků a výstupů projektu aplikační sférou a v rámci regionu

Prague Dependency Treebank, verze 1.0, publikováno v roce 2001:

Byly podepsány licence o výzkumném využití s 85 uživateli (62 organizací), z toho 19 uživatelů v ČR a SR, 27 v Evropě, 30 v USA, 7 v Asii, 2 na Stř. východě. Mezi nejvýznamnější zahraniční uživatele patří:

Z akademické nebo vládní oblasti:

Universita Komenského, Bratislava, Slovensko
University of Pennsylvannia, PA, USA
Brown University, RI, USA
Carnegie Melon University, PA, USA
Johns Hopkins University, MD, USA
The MITRE Corporation, Cambridge, MA, USA
New York University, NY, USA
U.S. Department of Defense, MD, USA
Tsinghua University, Peking, Čína
University of Stuttgart, Německo
Swiss Federal Institute of Technology (EPFL), Švýcarsko
Jozef Stefan Institute, Ljubljana, Zagreb, Chorvatsko
Univ. Of Szeged, Szeged, Maďarsko
Bulgarian Academy of Sciences, Sofia, Bulharsko

Řada licencí pochází i z průmyslových výzkumných center, např.:

AT&T, Florham Park, NJ, USA
Siemens AG, Německo
Lucent Technologies, Murray Hill, NJ, USA
NTT Communication Science Laboratories, Japonsko
Morphologic, Inc., Maďarsko

5. Podpora a výchova mladých výzkumných pracovníků

a) Doktorské studijní programy

Dvě sekce CKL při univerzitách (MFF UK a ZČU) jsou významným způsobem zapojeny do programů Doktorského studia. Podílejí se na výchově mladých výzkumných pracovníků, jimž jednak poskytují příležitost k vlastnímu bádání a k jeho prezentaci, jednak umožňují jejich zapojení do větších výzkumných úkolů.
Řada pracovníků Centra se podílí na výuce v rámci magisterských i doktorských studijních programů. V letním semestru 2002/2003 vedli celkem 33 přednášek a seminářů na MFF UK, FF UK a KK ZČU. V zimním semestru 2003/2004 vedli 28 přednášek a seminářů na MFF UK, FF UK a KK ZČU. Mimo to byli vedoucími řady studentských projektů.
Do doktorských studijních programů byli začátkem akademického roku 2003-2004 zapojeni další studenti, kteří jsou školeni pracovníky Centra; témata jejich disertačních prací jsou součástí vědeckého programu Centra.

školitel	# doktorandů
prof. Hajičová	7
doc. Hajič	16
dr. Kuboň	2
prof. Panevová	5
prof. Psutka	5
dr. Vidová-Hladká	2

Pracovník Centra Ing. Pavel Ircing (školitel prof. Psutka) odevzdal v září 2003 svoji disertaci na téma „Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language (Czech)". Disertace je v současné době v oponentním řízení.
Dále pět doktorandů zaměstnaných v CKL úspěšně složilo doktorské zkoušky, které jsou nezbytným předpokladem pro podání disertační práce.

V roce 2003 pokračovala spolupráce CKL s řadou nadaných studentů magisterského studia (a to jak z matematicko fyzikální fakulty a z filozofické fakulty UK, tak z katedry kybernetiky ZČU), u nichž po dokončení magisterského studia připadá v úvahu doktorandské studium. Většinou vykonávali pomocné práce týkající se anotací korpusových textů na různých úrovních, anotací a transkripcí řečových nahrávek, přípravy slovníků ap. Za svou práci byli odměňováni stipendiem, příp. měli uzavřený částečný úvazek (viz bod 2. Personální a organizační zabezpečení činnosti Centra).

Přístup k odborným časopisům a publikacím umožňuje všem výzkumným pracovníkům sledovat aktuální vývoj na předních zahraničních pracovištích oboru. Dobré přístrojové vybavení dovoluje efektivní práci na vlastních tématech.

CKL umožňuje v co nejširší míře svým pracovníkům prezentaci výsledků na mezinárodních konferencích a jejich konfrontaci s přístupy k podobným problémům ve světě. Podporuje účast mladých pracovníků na mezinárodních letních školách i jejich pracovní pobyty na zahraničních pracovištích (konkrétní akce jsou uvedeny v bodu 6.).

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

Jak bylo konstatováno v bodu 2. Personální a organizační zabezpečení činnosti Centra, mladí vědečtí a výzkumní pracovníci nadpoloviční část pracovníků CKL – 25 pracovníků CKL ze 40 je mladších než 35 let, dohromady zastávají 17,1 plných úvazků 24,7 úvazků (tj. více než 69%). Další objem práce odvádějí studenti magisterského studia (všichni do 35 let).

Lze tedy říci, že podpora mladých pracovníků Centra je velmi intenzívní. Vedle pravidelných

seminářů, na nichž vystupují se svými referáty, podíleli se mladí pracovníci podstatnou měrou i na lednovém a zářijovém výjezdním semináři (viz bod 1. Další aktivity Centra v roce 2003). 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 svědčí i počet přijatých referátů na mezinárodních konferencích; účast mladých pracovníků na těchto konferencích je díky podpory projektu velmi hojná a aktivní (konkrétní příklady viz bod 6.).

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

Účast studentů a doktorandů na mezinárodních konferencích v zahraničí

Zde uvádíme konkrétní zahraniční cesty mladých výzkumných pracovníků, které Centrum umožnilo:

Ondřej Bojar

- Vídeň, Rakousko, srpen 2003

Účast na letní škole ESSLLI 2003; poster o výsledcích automatické extrakce údajů z korpusu.

Alena Böhmová

- Paříž, Francie, srpen 2003

Účast na konferenci – projekt ENABLER.

Jan Cuřín

- Budapešť, Maďarsko, duben 2003

Účast na konferenci EACL 2003, spoluautor příspěvku na hlavní konferenci *Czech-English Dependency-based Machine Translation*

- Edmonton, Kanada, květen 2003

Účast na mezinárodní konferenci HLT-NAACL 2003.

- Växjö, Švédsko, listopad 2003

Účast na workshopu TLT2003, poster s názvem *Treebank in Machine Translation.*

Martin Čmejrek

- Budapešť, Maďarsko, duben 2003

Účast na konferenci EACL 2003, přednesení příspěvku na hlavní konferenci *Czech-English Dependency-based Machine Translation*.

- New Orleans, USA, září 2003

Účast na konferenci MT Summit 03 věnované strojovému překladu; účast na workshopech věnovaných evaluaci strojového překladu a jeho výuce.

- Växjö, Švédsko, listopad 2003

Účast na workshopu TLT2003, poster s názvem *Treebank in Machine Translation.*

Jiří Havelka

- Budapešť, Maďarsko, duben 2003

Účast na konferenci EACL 2003, spoluautor referátu na hlavní konferenci.

- Vídeň, Rakousko, srpen 2003

Účast na letní škole ESSLLI 2003.

Martin Holub

- Shefield, Anglie, květen 2003

Přípravná jednání pro letní workshop, pracovní schůzka.

- Řím, Itálie, červen 2003

Přípravné jednání – program na letní workshop v Baltimore.

- Baltimore, USA, červenec 2003

Účast na letní škole Johns Hopkins University Summer School on Human Language Technology.

Účast na mezinárodním výzkumném projektu v CLSP – CLSP Workshop 2003: Semantic Analysis Over Sparse Data.

- Dublin, Irsko, září 2003

Účast na konferenci International Symposium on Information and Communication Technologies 2003, příspěvek publikovaný ve sborníku *A New Approach to Conceptual Document Indexing: Building a Hierarchical System of Concepts Based on Document Clusters.*

Petr Homola

- Vídeň, Rakousko, srpen 2003

Účast na letní škole ESSLLI 2003.

- Sofie, Bulharsko, září 2003

Účast na konferenci Recent advances in NLP 2003, přednesen referát *Adapting SProUT to processing Baltic and Slavonic languages*.

- Tbilisi, Gruzie, říjen 2003

Účast na konferenci 5th International Symposium on Logic, Language and Computation, prezentace experimentálního systém strojového překladu z češtiny do litevštiny *Shallow machine translation - in between of two extremes.*

Veronika Kolářová-Řezníčková

- Krakow, Polsko, březen 2003

Účast na intenzivních pracovních schůzkách s prof. Stanislavem Karolakem, věnovaných otázkám syntaxe slovanských jazyků ve 2. pol. 20. století. Pozornost byla zaměřena zejména na možnosti zachycení predikáto-argumentových struktur sloves získaných z různých typů korpusů.

- Lancaster, Anglie, březen 2003

Aktivní účast na konfernci Corpus Linguistics 2003. Na workshopu "The Shallow Processing of Large Corpora Workshop (SProLaC 2003)" přednesen referát *Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora.*

- Bratislava, Slovensko, listopad 2003

Účast na konferenci Slovanské jazyky v počítačovom spracovaniu (SLOVKO 03), přednesen referát *Využití ČNK a PZK pro ověřování valenčních vlastností některých typů deverbativních substantiv.*

- Prešov, Slovensko, prosinec 2003

Seznamování budoucích anotátorů Slovenského národního korpusu s anotačními nástroji vyvinutými v CKL a prezentace syntaktického značkování na tektogramatické rovině. Dále konzultace věnované valenčnímu slovníku českých sloves VALLEX.

Jiří Kocanda

- Baltimore, USA, červenec 2003

Účast na letní škole Johns Hopkins University Summer School on Human Language Technology.

Pracovní návštěva LDC, University of Pennsylvania, Philadelphia.

Kateřina Marková

- Paříž, Francie, červen 2003

Účast na konferenci MIT 2003 (první mezinárodní konference o Meaning Text Theory).

Jiří Mírovský

- Philadelphia, Pennsylvania, USA, listopad 2003

Pracovní návštěva LDC (University of Pennsylvania), účast na dvoutýdenním workshopu zaměřeném na vyhledávání v arabském Treebanku. Hlavním cílem cesty bylo zprovoznění vyhledávacího nástroje Netgraph, přednáška na toto téma a školení.

J. Mlejnecká

- Tubingen, 15.-20.9.2003

Konference Polyslav, referát *Stereotypy v jazyce*.

- **Berlín, 2003**

Veletrh Expolingua.

Roman Ondruška

- Lancaster, Anglie, březen 2003

Aktivní účast na konferenci Corpus Linguistics 2003, referát *An Exploitation of the Prague Dependency Treebank: A Valency Case.*

- Philadelphia, Pennsylvania, USA, listopad 2003

Pracovní návštěva LDC (University of Pennsylvania), účast na dvoutýdenním workshopu zaměřeném na vyhledávání v arabském Treebanku. Hlavním cílem cesty bylo zprovoznění vyhledávacího nástroje Netgraph, přednáška na toto téma a školení.

Nino Peterek

- Ženeva, Švýcarsko, září 2003

Účast na konferenci EUROSPEECH 2003.

Petr Podveský

- Magdeburg, Německo, červenec 2003

Účast na letní škole o rozpoznávání řeči.

- Ženeva, Švýcarsko, září 2003

Účast na konferenci Eurospeech 2003.

Kiril Ribarov

- Lublaň, Slovinsko, srpen 2003

Účast na 13. Mezinárodním kongresu slavistů; účast na schůzi komise pro počítačové zpracování starých památek a prvotisků, zvolen jejím předsedou.

Karolina Skwarska

- Lublaň, Slovinsko, srpen 2003

Účast na 13. Mezinárodním kongresu slavistů.

Otakar Smrž

- Budapešť, Maďarsko, duben 2003

Účast na konferenci EACL 2003, přednesení referátu v rámci sekce Research Notes *Arabic Syntactic Trees: from Constituency to Dependency* (spoluautor Z. Žabokrtský).

- Baltimore, USA, červenec 2003

Účast na letní škole Johns Hopkins University Summer School on Human Language Technology.

Pracovní návštěva LDC, University of Pennsylvania, Philadelphia.

K. Smejkalová

- Tübingen, 15.- 20.9. 2003

Konference Polyslav, referát *Negace v přísudku české věty (na základě materiálu ČNK, Treebanku a poradenské databáze).*

Jan Štěpánek

- Lancaster, Anglie, březen 2003

Aktivní účast na konfernci Corpus Linguistics 2003, přednesení referátu *An Exploitation of the Prague Dependency Treebank: A Valency Case.*

- Vídeň, Rakousko, srpen 2003

Účast na letní škole ESSLLI 2003.

Zdeněk Žabokrtský

- Saarbrücken, Německo, březen 2003

- Vylepšování softwarových nástrojů pro automatický převod německých vět z korpusu Negra do tektogramatické roviny Pražského závislostního korpusu.
- Návrh a implementace rozhraní pro anotaci gramatémů v českých i německých tektogramatických stromech.
- Zahájení spolupráce na téma anotování koreferenčních vztahů.

- Budapešť, Maďarsko, duben 2003

Účast na konferenci EACL 2003, spoluautor referátu *Arabic Syntactic Trees: from Constituency to Dependency.*

- Lublaň, Slovinsko, květen 2003

Ve spolupráci mezi Institutem Jozefa Štefana a Filozofickou fakultou v Lublani zahájení budování syntakticky anotovaného korpusu slovinštiny. Tento projekt bude v maximální možné míře využívat závislostního popisu syntaxe i softwarových nástrojů vyvinutých v rámci projektu Pražský závislostní korpus, což bylo hlavním cílem cesty.

- Brighton, Velká Británie, prosinec 2003

Účast na workshopu 2nd International Workshop on Dictionary Writing Systems.

7. Způsoby zpřístupnění výsledků a výstupů Centra veřejnosti

Prezentace výsledků Centra na konferencích

Pracovníci CKL se zúčastnili řady mezinárodních konferencí a jiných odborných setkání, na kterých přednesli zvané přednášky (12 zvaných přednášek) a recenzované příspěvky o výsledcích dosažených v projektech CKL, případně prezentovali své výsledky na posterech. Participovali na workshopech, které jsou vynikající příležitostí pro sdílení nových výsledků a postupů.

Byl zorganizován workshop o diskurzu při Mezinárodním kongresu lingvistů.

K prestižním konferencím a workshopům, které mají velký ohlas mezi odbornou veřejností, patřily zejména následující:

- mezinárodní konference **Corpus Linguistics 2003** v Lancasteru, březen 2003 (2 vystoupení)
- mezinárodní konference **EACL 2003** v Budapešti, březen 2003 (2 vystoupení)
- mezinárodní **Workshopu SSPR 2003** při konferenci ISCA&IEEE v Tokiu, duben 2003 (1 vystoupení)
- mezinárodní konference **MIT 2003** v Paříži, červen 2003 (1 vystoupení)
- **Světový kongres lingvistů** v Praze, červenec 2003 (7 vystoupení + 2 postery)
- **Mezinárodní kongres slavistů** v Lublani, srpen 2003 (2 vystoupení)
- mezinárodní konference **Eurospeech 2003** v Ženevě, září 2003 (2 vystoupení)
- mezinárodní konference **TSD 2003** v Českých Budějovicích, září 2003 (6 vystoupení)
- mezinárodní konference **Machine Translation Summit 03** a přidružené workshopy, New Orleans, září 2003 (1 vystoupení)
- mezinárodní symposium bohemistů **Čeština – univerzália a specifika**, Brno, listopad 2003 (4 vystoupení + 1 poster)
- mezinárodní Workshop on Treebanks and Linguistic Theories, Vaxjo, listopad 2003 (2 vystoupení + 1 poster)

Publikace

Publikace v domácích i zahraničních časopisech a ve sbornících mezinárodních konferencí zpřístupňují výsledky Centra široké odborné veřejnosti. Celkem bylo připraveno 68 publikací (seznam viz Přílohu 2).

Technické zprávy

CKL **pokračovalo** ve vydávání **technických zpráv** (ve spolupráci s ÚFALem MFF UK) o dílčích výsledcích výzkumu; v roce 2003 byly vydány 4 výzkumné zprávy, které jsou též k dispozici na webových stránkách CKL http://ckl.mff.cuni.cz:8080/pub/publications.jsp?type=tr.

- Květoň, P. (2003) Language for Grammatical Rules. CKL/UFAL Technical Report TR-2003-17.
- Lopatková, M., Žabokrtský, Z., Skwarska, K., Benešová, V. (2003) Valency Lexicon of Czech Verbs VALLEX 1.0. CKL/UFAL Technical Report TR-2003-18.
- Kučová, L., Kolářová, V., Pajas, P., Žabokrtský, Z., Culo, O. (2003) Anotování koreference v Pražském závislostním korpusu. CKL/UFAL Technical Report TR-2003-19.
- Veselá, K., Havelka, J. (2003) Anotování aktuálního členění věty v Pražském závislostním korpusu. CKL/UFAL Technical Report TR-2003-20

Webové stránky CKL

Byla dokončena restrukturalizace www stránek CKL, http://ckl.mff.cuni.cz/, která začala v předchozím období. V současné době stránky podávají komplexní informaci o činnosti Centra. Kromě základních informací o struktuře, výzkumných tématech a cílech CKL zde lze nalézt zejména odkazy na stránky jednotlivých úkolů řešených v rámci CKL (PDT, ČAK, VALLEX, PADT, MALACH, VMC, CILXVII, viz bod 1. Stručný přehled dílčích cílů projektu), včetně volně šiřitelných nástrojů (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.

ENABLER Network

CKL jako člen mezinárodní sítě ENABLER Network (zaměřené na shromažďování informací o existujících jazykových zdrojích pro jednotlivé jazyky) průběžně aktualizuje přehled dostupných jazykových zdrojů, které vznikají v rámci jednotlivých projektů Centra, čímž zásadně přispívá 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 (viz Přílohu 1.)

Den otevřených dveří

Dne 2.12. v rámci Dnu otevřených dveří na MFF UK byla širší odborné veřejnosti živou formou představena činnost CKL a zájemci především ze středních škol byli seznámeni s tématy, na nichž pracujeme, a také s našimi výsledky.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1,110 00, IČO: 00216208

Podrobná specifikace a zdůvodnění jednotlivých položek finančních prostředků projektu čerpaných v r. 2003

1. *Rozpis celkových výdajů ve sledovaném období za všechny účastníky projektu, (tj. vklady zakládajících organizací + účelové prostředky požadované na zadavateli, v tisících)*

Prostředky	**Centrum**	
Investice	**447***	**(+3)**
Neinvestice	**28 739**	**(- 3)**
Mzdy	10 137	(0)
z toho OON	550	(0)
Režie	6 996	(0)
Ostatní	11 606	(- 3)
Ostatní podrobně		
Odpisy	2 988	(0)
Cestovné a pobyty	2 766	(+5)
Pojištění	3 477**	(+71)
DHM a NHM	785	(- 18)
Materiál	488	(+2)
Služby	303**	(- 77)
Další	799	(+14)
Celkem	**29 186**	**(0)**

* Pro rok 2003 byla CKL přidělena na investice pouze částka 450 tisíc. Částka 2 100 tisíc na investice na rok 2003 byla na základě písemné žádosti sekce CKL na MFF UK se souhlasem zadavatele převedena do roku 2002, kdy byly prostředky vyčerpány.

Komentář:
** Částka 70 tis. byla se souhlasem zadavatele převedena z kolonky *Pojištění* do kolonky *Služby*. Z položky OON byly hrazeny krátkodobé práce na základě dohody o provedení práce, čímž byla položka na pojištění nižší.

Poznámka:
Tabulka odpovídá kalkulaci předpokládaných finančních výdajů v r. 2003, jak byla specifikována ve zprávě Centra za rok 2002 a v kalkulaci ve Zdůvodnění návrhu projektu. Byla opravena výpočetní chyba v položce *Ostatní* (chybný součet v kolonce *další*).

2. *Specifikace a zdůvodnění jednotlivých výdajových položek ve vztahu k projektu,*

MFF UK

Investice viz bod 3.

Neinvestice (plán 16 782 tis.)

Mzdy (plán 8 689 tis.)
Mzdové prostředky byly podle plánu využity na platy a odměny zaměstnanců, viz bod 2. Personální a organizační zabezpečení Centra.

OON (plán 550 tis.)
Prostředky využity podle plánu.

Režie (plán 1 349 tis.)
Prostředky na režii využity podle plánu.

Ostatní (plán 6 744 tis.)

Cestovné a pobyty (plán 2 398 tis.)
Cestovné bylo podle plánu použito na úhradu cest pracovníků CKL na zahraniční i domácí konference, kde prezentovali své výsledky (viz seznam zahraničních cest, bod 1.)

Pojištění (plán 3 041 tis.)
Částka 70 tis. byla se souhlasem zadavatele převedena do kolonky *Služby*. Z položky OON byly hrazeny krátkodobé práce na základě dohody o provedení práce, čímž byla položka na pojištění nižší.

DHM a NHM (plán 532 tis.)
Prostředky využity podle plánu.

Materiál (plán 415 tis.)
Prostředky využity podle plánu.

Služby (plán 67 tis.)
Položka *Služby* byla se souhlasem zadavatele navýšena o částku 70 tis. z položky *Pojištění*, ze které byly hrazeny vyšší náklady na publikace.

Další (plán 290 tis.)
Prostředky využity podle plánu.

ZČU

Investice – viz bod 3.

Neinvestice (přiděleno 1 672 tis. Kč)

Mzdy (plán 548 tis. Kč)
Za měsíce leden až prosinec 2003 bylo vyplaceno 548 tis. Kč

OON (plán 0 Kč)
Za měsíce leden až prosinec 2003 bylo vyplaceno 0 Kč

Režie (plán 187 tis. Kč)
Režie 187 tis. Kč

Ostatní (plán **937 tis**. Kč)

Cestovné (plán 213 tis. Kč)
Konference Eurospeech 2003, TSD 2003, SSPR 2003, SCI 2003 198 tis. Kč
Cestovné tuzemské 10,3 tis. Kč
Celkem 208,3 tis. Kč

Pojištění sociální a zdravotní (plán 192 tis. Kč)
Pojištění 188 tis. Kč

DHM a NHM (plán 101 tis. Kč)
Přenosné disky 25,4 tis. Kč

Řečové korpusy		15,2 tis. Kč
Notebook		39,1 tis. Kč
DVD vypalovačka		17,0 tis. Kč
Knihy		9,7 tis. Kč
Řadič		1,8 tis. Kč
	Celkem	108,2 tis. Kč
Materiál (plán 32 tis. Kč)		
Kancelářský materiál		37,3 tis. Kč
DVD		3,0 tis. Kč
	Celkem	40,3 tis. Kč
Služby (plán 133 tis. Kč)		
Vložné TSD 2003		22,1 tis. Kč
Vložné na konf. (SCI2003, Eurospeech2003, SSP2003)		69,3 tis. Kč
Vnitro-služby		28,7 tis. Kč
Ostatní služby (bankovní, telefonní, poštovné, přepravné ap.)		19,3 tis. Kč
	Celkem	139,4 tis. Kč
Další (plán 266 tis. Kč)		
Místnosti (teplo, energie ap.)		100 tis. Kč
Stipendia		100 tis. Kč
Oprava a údržba		42,9 tis. Kč
Tiskové služby		11,4 tis. Kč
	Celkem	254,3 tis. Kč

Neinvestice celkově (k 8.1.04) **1 672 000 Kč**

ÚJČ

Investice viz bod 3.

Neinvestice (plán 1 615 tis.)

Mzdy (plán 665 tis.)
Mzdové prostředky byly využity na platy a odměny zaměstnanů, viz bod 2. Personální a organizační zabezpečení Centra.

Režie (plán 158 tis.)
Prostředky využity podle plánu.

Ostatní (plán 792 tis.)

Cestovné a pobyty
Cestovné bylo použito na úhradu cest pracovníků CKL na zahraniční i domácí konference, kde prezentovali své výsledky (viz seznam zahraničních cest, bod 1.)

Pojištění
Prostředky využity podle plánu.

DHM a NHM
Mimo jiné byly zakoupeny tři ploché obrazovky (2x 17 “ a 1x 19 “).

Materiál
Prostředky využity podle plánu.

Služby
Prostředky využity podle plánu.

Další
Prostředky v položce *Další* byly využity podle plánu na služby (tvorba databází pro ukládání korpusových nálezů), tisk posteru a další.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1,110 00, IČO: 00216208

3. Jednoznačná specifikace položek hrazených z účelové dotace

Plán:

Účelové prostředky	**Centrum**	**MFF UK**	**ZČU**	**ÚJČ**
Investice	**450**	**0**	**300**	**150**
Neinvestice	**20 070**	**16 782**	**1 672**	**1 615**
Mzdy	9 902	8 689	548	665
z toho OON	550	550	0	0
Režie	1 695	1 349	187	158
Ostatní	8 473	6 744	937	792
Ostatní podrobně				
Cestovné a pobyty	2 771	2 398	213	160
Pojištění	3 466	2 971*	192	233
DHM a NHM	767	532	101	133
Materiál	490	415	32	43
Služby	226	137*	133	27
Další	753	290	266	197
Celkem	**20 520**	**16 782**	**1 972**	**1 765**

Komentář:
* Částka 70 tis. byla se souhlasem zadavatele převedena z kolonky *Pojištění* do kolonky *Služby*. Z položky OON byly hrazeny krátkodobé práce na základě dohody o provedení práce, čímž byla položka na pojištění nižší.

Skutečnost:

Účelové prostředky	**Centrum**	**MFF UK**	**ZČU**	**ÚJČ**
Investice	**447 (+3)**	**0 (0)**	**300 (0)**	**147 (+3)**
Neinvestice	**20 073 (- 3)**	**16 782** (0)	**1 672** (0)	**1 618 (- 3)**
Mzdy	9 902 (0)	8 689 (0)	548 (0)	665 (0)
z toho OON	550 (0)	550 (0)	0 (0)	0 (0)
Režie	1 695 (0)	1 349 (0)	187 (0)	158 (0)
Ostatní	8 476 (-3)	6 744 (0)	937 (0)	795 (- 3)
Ostatní podrobně				
Cestovné a pobyty	2 766 (+5)	2 398 (0)	208 (+5)	160 (0)
Pojištění	3 395 (+71)	2 972 (- 1)	188 (+4)	235 (- 2)
DHM a NHM	785 (-18)	532 (0)	108 (-7)	145 (-12)
Materiál	488 (+ 2)	416 (- 1)	40 (-8)	32 (+11)
Služby	303 (-77)	137 (0)	139 (-6)	27 (0)
Další	739 (+14)	288 (+2)	254 (+12)	197 (0)
Celkem	**20 520 (0)**	**16 782 (0)**	**1 972 (0)**	**1 765 (0)**

Komentář:
V závorkách jsou uvedena čísla, která odpovídají případnému přečerpání (-) / nedočerpání (+) prostředků.

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

Pokud došlo ve sledovaném období ke změnám ve specifikaci jednotlivých smlouvou stanovených finančních položek, je nutno uvést stručnou charakteristiku těchto změn v porovnání se smlouvou, zdůvodnit je ve vztahu k cílům projektu a uvést stanovisko poskytovatele.

Prostředky nositele	Celkem	MFF UK	ZČU	ÚJČ
Investice	**0**	**0**	**0**	**0**
Neinvestice	**8 666**	**7 369**	**444**	**853**
Mzdy	235	157	0	78
Režie	5 301	4 481	201	619
Odpisy	2 988	2 657	213	119
Místnosti	20	0	20	0
Pojištění	82	55	0	27
Další	40	20	10	10
Celkem	**8 666**	**7 369**	**444**	**853**

V roce 2000 byly investice hrazeny (nakupovány) na všech pracovištích CKL, a na jednotlivých pracovištích jsou také odpisovány. V roce 2001 byly veškeré investice pořizovány na pracovišti nositele, tj. MFF UK (a na pracoviště spolupříjemců byly zapůjčeny), proto jsou investice z r. 2001 odpisovány pouze na MFF (na MFF byly tedy odpisy zvýšeny a na spoluřešitelských pracovištích odpovídajícím způsobem sníženy). V roce 2002 a 2003 opět jednotlivá pracoviště realizovala investice samostatně, a také je sama odepisují.
Tomuto schématu odpovídá i uvedená tabulka:

- výše odpisů na MFF UK byla zvýšena o odpisy z investic za rok 2001, které podle původního plánu měly být majetkem ZČU a ÚJČ
- naopak výše odpisů na ZČU a ÚJČ byla o odpovídající částku snížena

Komentář: Vklad ZČU byl realizován na účet projektu ve výši 213 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 a v roce 2003 (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.

Poznámka:
Položky a částky odpovídají tabulce ve formátu *. xls.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1,110 00, IČO: 00216208

Přehled a upřesnění dílčích cílů projektu a postupu při jejich naplňování pro následující období, tj. pro r. 2004[2]

(Specifikace dílčích cílů pro následující období musí být v souladu s cíli stanovenými smlouvou!)

Uvádíme zde upřesnění cílů pro r. 2004, jak byly stanoveny v šesti bodech (A) až (F) v původním návrhu programu. Cíle jsou konkretizovány v rámci tří nosných výzkumných projektů, které vykrystalizovaly v prvních 18 měsících existence Centra a byly též takto prezentovány ve výročních zprávách za léta 2002 a 2003. Jsou to: rozvíjení Pražského závislostního korpusu (bod B původního návrhu), projekt strojového překladu (bod F původního návrhu) a v rámci výzkumu v oblasti zpracování mluvené řeči pak participace na mimořádně rozsáhlém mezinárodním projektu MALACH (bod E původního návrhu). Souběžně s těmito projekty a v návaznosti na ně bude pokračovat výzkum v oblasti teoretických aspektů komputační lingvistiky, tedy jejích matematických i lingvistických základů (body A, C a D) a rovněž vyvíjení některých aplikačních systémů (bod F původního návrhu).

V následujícím přehledu jsou jednotlivé body konkretizovaného programu pro rok 2004 označeny (T1) až (T6) , popř. dalším členěním, a to v souladu s časovým harmonogramem uvedeným pod přehledem.

T1: Rozvíjení Pražského závislostního korpusu

Pražský závislostní korpus (PDT) je stěžejním projektem CKL. Výzkum v roce 2004 bude plynule navazovat na výsledky let předcházejících, v jistém smyslu bude dovršením první etapy vývoje PDT a bude vytvářet předpoklady jak pro další etapu jeho rozvoje, tak pro jeho využití uživateli z nejrůznějších oblastí výzkumné i aplikační činnosti. Pozornost se soustředí především na následující body.

T1-1 Výsledná lexikálně-morfologická databáze bude zveřejněna pomocí online přístupu, spolu se systémem registrace přístupu a zadání uživatelů, z něhož bude možno v budoucnu (případném pokračování projektu Centra) doplňovat tuto databázi.
V oblasti syntaktické budou zveřejněny lexikální zdroje použité pro budování Pražského závislostního korpusu, a to rovněž formou online přístupu (slovníky PDT-VALLEX a VALLEX, již částečně zveřejněný v r. 2003).

T1-2 Navržený algoritmus pro určení valence deverbativ bude verifikován na malém úseku korpusu.

T1-3 Anotace základních rysů aktuálního členění bude přiřazena všem větám ve velkém souboru. Na menším vzorku takto označených vět bude ověřen algoritmus určení základu a ohniska věty.

T1-4 Budou zjišťovány možnosti přechodu od tektogramatických struktur s vyznačením aktuálního členění a základních rysů koreference k zachycení struktury diskurzu založenému na stupních aktivace ve sdílené zásobě znalostí.

T1-5 Bude aplikovány principy rekonstrukce uzlů v tektogramatickém stromě a upřesněna specifikace lexikálních lemat pro tyto uzly.

T1-6,7 Bude verifikována automatická procedura při anotaci velkého korpusu.

T1-8 Budou konfrontovány možnosti a výsledky vyhledávek v ČNK s možnostmi a výsledky vyhledávek v Pražském závislostním korpusu. Vybrané výsledky těchto konfrontací budou

[2] Uvádí se bližší specifikace cílů stanovených smlouvou a jejich rozpis na dílčích cíle pro daný kalendářní rok, vč. časového harmonogramu

ukládány ve speciálních databázích, které budou k tomu účelu budovány jako složky internetových stránek.

T1-9 Bude dokončena konverze vnitřního kódování Českého anotovaného korpusu a následně bude zveřejněna.

Hlavním výsledkem bodu T1 a do jisté míry i vizitkou celého Centra komputační lingvistiky bude zveřejnění (publikace) CD ROM s Pražským závislostním koprusem verze 2.0, která bude proti verzi 1.0 (2001) obsahovat navíc 55 tisíc vět anotovaných velmi podrobně na tzv. tektogramatické rovině, a řadu nových a vylepšených nástrojů na zpracování češtiny. S finální publikací (opět prostřednictvím Linguistic Data Consortium, Univ. of Pennsylvannia, PA, USA) se počítá v závěru roku 2004; do té doby bude probíhat anotace dosud neanotovaných jevů, oprava inkonzistencí, sjednocování formátu, tvorba finální dokumentace, a zejména kontorly specifikace a dokumentace vůči datům.

T2: Strojový překlad

V oblasti strojového překladu se zaměříme na statisticky založený systém anglicko-český a na další experimenty s překladem mezi jazyky blízkými a jejich shluky se společným pivotním jazykem (češtinou). Další zdroje budou věnovány na vytvoření veřejně dostupných datových zdrojů a nástrojů pro strojový překlad mezi češtinou a angličtinou, a to vzhledem k velké poptávce po takových zdrojích a k oživení v oblasti základního i aplikovaného výzkumu v oblasti strojového překladu v souvislosti se vstupem do EU, a s ohledem na budoucí priority národního programu výzkumu v oblasti překonávání jazykových bariér. Vytvořené zdroje a nástroje budou završením práce Centra na zdrojích a nástrojích pro strojový překlad a východiskem pro další práci v případném budoucím pokračování projektu. pozornost bude soustředěna zejména na:

T2-1 Shrnutí výsledků experimentů se statistickým překladem založeným na závislostní struktuře.

T2-2 Experimenty ověřující možné použití automatického překladu nominálních frází lidským překladatelem. Pokračování v integraci dalších lingvistických zdrojů (Penn Treebank, PropBank) do překladového systému.

T2-3 Experimenty s interpolací skóre překladového a jazykového modelu při automatickém rozpoznávání češtiny. Začlenění slovníku do CD připravovaného k vydání v LDC.

T2-4 Vydání paralelního korpusu a programových nástrojů pro česko-anglický strojový překlad na CD-ROM v Linguistic Data Consortium.

T3: Zpracování mluvené řeči

T3-1 V roce 2004 bude dále pokračovat především řešení projektu MALACH. V systému pro rozpoznávání češtiny se bude řešit problematika nespisovných slov, případně výskytu gramaticky nekorektních vazeb, a budou provedeny též experimenty s jazykovými modely založenými na morfologických značkách.
Pro ruštinu bude natrénován nový systém rozpoznávání spontánní řeči. Stávající akustické a jazykové modely byly zatím natrénovány pouze na malém množství dat a sloužily především k ověření správnosti navržené fonetické abecedy. Pozornost bude věnována také specifickým problémům, které při rozpoznávání rusky mluvených svědectví nastanou. Jako hlavní problém se v současné době jeví nestandardní výslovnost některých slov mluvčími, kteří žili dlouhou dobu mimo území Ruska (ze 7 tisíc výpovědí vedených v ruském jazyce bylo pouze asi 700 poskytnuto na území Ruska - ostatní výpovědi poskytli lidé žijící většinou již dlouhou dobu na území mimo Ruska, např. na Ukrajině, v Izraeli, v USA ap.).
Dále probíhá anotace slovenských výpovědí, jejichž rozpoznávání je dalším úkolem v příští fázi projektu - „ověřovací“ systém bude vyvinut v průběhu roku 2004. Budou též připravovány nástroje a specifikován postup pro anotační práce pro zpracování výpovědí vedených v dalším

jazyku - mluvené spontánní polštině.
V roce 2004 budou pokračovat práce i na úloze automatického přepisu komentářů hokejových utkání. Dlouhodobým cílem této úlohy je automatické titulkování komentářů, které by zabezpečilo nikoli přesný přepis (to zatím současné světové technologie nezvládají), ale přepis, který by byl gramaticky korektní a nesl by pokud možno shodnou sémantickou informaci s mluveným komentářem.

T3-2 Budou analyzovány F0 křivky a dále bude ověřován vlivu rychlosti a energie promluvy na prozodické opozice.

T3-3 Budou pokračovat rovněž práce na digitalizaci dat AV ÚJČ a do již přepsaných dokumentů budou vloženy časové značky, které umožní detailní prozodickou analýzu. Dosud nashromážděná data budou použita při vývoji a ověřování automatických značkovacích nástrojů prozodie. Pokračovat se bude ve studiu prozodických opozic mezi základem, kontrastivním základem, průvodními členy základu a vlastním ohniskem na základě rozsáhlých dat a vizualizace prozodických parametrů.

T4: Teoretické aspekty komputační lingvistiky, její matematické i lingvistické základy (body A, C a D původního návrhu)
Teoretický výzkum v rámci Centra je neoddělitelně spjat s výše zmíněnými projekty, a to jednak jako předpoklad pro jejich formulaci a teoretický základ pro jejich řešení, jednak tyto projekty přinášejí vedle ověřování platnosti navržených hypotéz i důležité další podněty pro teoretické bádání. V roce 2004 bude výzkum pokračovat v následujících bodech:

T4-1 Na základě výsledků experimentů dosažených v roce 2003 (Bod T4-1, přehled splněných cílů r. 2003) bylo rozhodnuto v tomto bodě dále nepokračovat. Další nutný teoretický výzkum příslušných metod bude zajištěn mimo CKL. Uvolněná finanční kapacita bude použita na krytí nových aktivit Centra, viz kap. „Nové projekty Centra“ a zejména metody a zdroje strojového překladu.

T4-2 V oblasti syntaktického zpracování (parsing) se bude nadále pracovat na možnostech kombinace parserů, a práce budou nadále probíhat i v oblasti zlepšování úspěšnosti existujících parserů. Dále bude připraven základní systém syntakticko-sémantického parsingu, který bude vycházet z povrchově-syntaktické reprezentace, jako výsledku projektu Centra v oblasti syntaktické analýzy. Systém bude vyhodnocen, a bude považován za tzv. baseline pro případné pokračování projektu.

T4-3 V souvislosti s budování zdrojů pro strojový překlad (T2) se budou rozvíjet jazykově nezávislé metody získávání slovníkových dat z korpusů, a to vícejazyčných i jednojazyčných. Bude se pokračovat pokračovat v obohacování slovníku automatickými metodami a bude vydán česko-anglický překladový slovník v LDC.

T4-4 Jako završení činnosti Centra v oblasti metodologie strojového překladu bude zprovozněn kompletní systém strojového překladu z češtiny do angličtiny, který bude používat strukturní, tzv. tektogramatickou reprezentaci (vyvíjenou rovněž v CKL) jako reprezentaci struktur vět v jádru systému překladu. Tento systém bude vyhodnocen vzhledem k existujícím systémům založeným na jiné (např. lineární) reprezentaci nebo vůči systémům vytvořeným jinou technologií.

T4-5 V oblasti lingvistických základů komputační lingvistiky se bude pokračovat ve studiu významové (tektogramatické) struktury věty na základě dat získaných anotací PDT, a to především těchto aspektů:
- studium dalších skupina sloves (z hlediska vztahu polysémie a valence);
- studium hranic větné a členské koordinace;
- zařazení nových typů syntaktických funkcí do funkčního generativního popisu a do jeho

valenční teorie;
- klasifikace českých zájmen a číslovek (z hlediska „hloubkových“ lexikálních jednotek;
- v oblasti aktuálního členění věty pokračování ve studiu kontrastu, a to z hlediska možnosti či nutnosti rozlišení kontrastivní části ohniska věty a průvodních členů ohniska, resp. ohniska kontrastivního a idantifikačního;
- v oblasti diskurzu prohloubení formulovaného algoritmu přiřazování stupňů aktivovanosti jednotlivým členům věty a jeho aplikace na anotované soubory PDT;
- v oblasti vztahů mezi rovinami syntaxe a morfématiky zejména soustavné studium odchylek povrchového slovosledu od tektogramatické projektivity stromu.

T4-6 Pokračování ve studiu a specifikaci reprezentace těch sémantických (kognitivních) aspektů, které přesahují jazykový význam, pro případné doplnění anotačního scénáře PDT o další úroveň.

T5: Vyvíjení některých menších aplikačních systémů (bod F)

T5-1 Budou vyvinuty softwarové nástroje pro tvorbu ucelených prototypů strojového překladu, s modulárním řešením pro snadnější provádění experimentů v této oblasti a pro případné budoucí začlenění do aplikací.
Budou vypracovány moduly pro předzpracování překladu mezi blízkými jazyky s ohledem na profesionální překladové systémy pracující s překladovou pamětí.
T5-2 Bude ověřena možnost využití systém SProUT pro morfologické předzpracování a parciální syntaktickou analýzu pro systémy pracující s překladovou pamětí a jeho rozšíření pro baltoslovanské jazyky vyvíjený na DFKI (Německé výzkumné centrum pro umělou inteligenci) v Saarbrückenu.
T5-3 Bude dokončen projekt "Sémantické modely textu" a výsledný model evaluován s využitím anotovaných dat..

T6: Přípravy národních a mezinárodních akcí

T6-1 Dokončení organizační přípravy 19. běhu mezinárodních cyklů Centra Viléma Mathesia a zajištění jejich plynulého průběhu (březen 2004). Vyhodnocení cyklu a jeho vyúčtování.

T6-2 Příprava workshopu (jako organizátor) **Prague-Penn Arabic Treebanking** na mezinárodní konferenci Language Resources and Evaluation Conference (LREC 2004), Lisabon, Portugalsko, květen 2004.

T6-3 Příprava workshopu (jako spoluorganizátoři) **Beyond Named Entity Recognition - Semantic Labelling for Natural Language Processing Tasks** na mezinárodní konferenci Language Resources and Evaluation Conference (LREC 2004), Lisabon, Portugalsko, květen 2004.

T6-4 Příprava workshopu (jako spoluorganizátoři) na mezinárodní konferenci North American Association for Computational Linguistics (Boston, Massachusetts) nazvaného **Frontiers in Corpus Annotation**, květen 2004.

T6-5 Centrum bude společně s LDC, PA, USA připravovat výzkumně-školící projekt pro lingvistické anotace. Společný workshop se uskuteční v USA v červnu/červenci 2004, krátký tutorial bude nabídnut též na konferenci ACL 2004 v Barceloně.

T7: Nové projekty

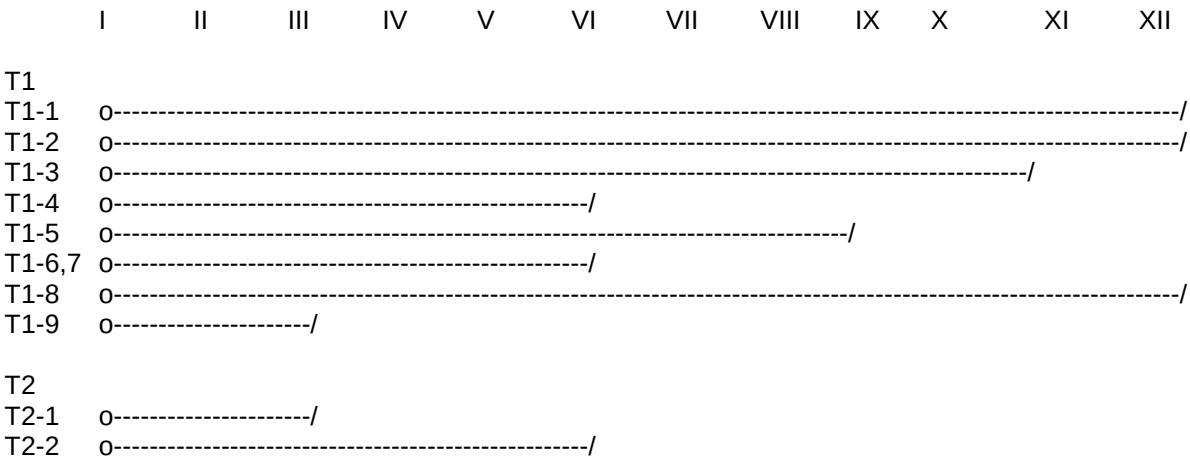
T7-1 Projekt **Prague Arabic Dependency Treebank** se soustředí na analytické značkování a na získávání podkladů pro tektogramatický popis arabské věty, aby bylo zajištěna možnost dalšího plynulého pokračování projektu.
T7-2 V rámci projektu **Zpracování psaného kulturního dědictví** bude vytvářen anotovaný korpus staroslověnštiny (naplňování databáze, cca 600 000 slovních forem), který bude zpřístupněn přes webové stránky CKL. Budou pokračovat práce na anotačních modulech systému ACT.
Bude odstartován projekt Transfer of Knowledge, kde pracoviště CKL má funkci "dodavatele" znalostí v oblasti anotování obecně a v oblasti zpracování psaného kulturního dědictví.

Další aktivity Centra plánované pro rok 2004

- Budeme **pokračovat** ve vydávání **technických zpráv** (ve spolupráci s ÚFALem MFF UK) o dílčích výsledcích výzkumu; v roce 2003 předpokládáme vydání nejméně tří výzkumných zpráv. Tyto zprávy budou též k dispozici na webových stránkách CKL.
- Alespoň na jednom z pracovišť Centra **uspořádáme Den otevřených dveří**, na němž seznámíme živou formou širší odbornou veřejnost a především zájemce ze středních škol s tématy, na nichž pracujeme, a s našimi výsledky.
- Předpokládáme krátkodobé příležitostné **přednáškové pobyty několika předních zahraničních profesorů**.
- Centrum se podstatnou měrou podílí a bude podílet na organizaci 19. běhu mezinárodních **cyklů přednášek Centra Viléma Mathesia** v Praze v březnu 2004. Kurzy budou vedeny 10 významnými zahraničními odborníky a nejméně čtyřmi pracovníky CKL. Veškeré organizační zajištění je dílem Centra. Bezplatně se přednášek zúčastní kolem 20 českých doktorandů a mladých vědeckých pracovníků.
- CKL uskuteční alespoň jeden **výjezdní seminář**, na kterém se budou soustavně projednávat jednotlivé úkoly Centra a mladí pracovníci Centra budou prezentovat své výsledky.

Poznámka k organizačnímu zabezpečení Centra
Vzhledem k tomu, že od 1. 1. 2004 dojde k rozdělení oddělení jazykové kultury a gramatiky na samostatné oddělení jazykové kultury (včetně jazykové poradny – vedoucí L. Uhlířová) a samostatné oddělení gramatiky (vedoucí Fr. Štícha), obě tato oddělení budou zapojena do činnosti CKL. Činnost oddělení gramatiky bude více zaměřena na využívání PDT při gramatickém výzkumu a na vyhledávání chybného tagování, oddělení jazykové kultury bude zaměřeno na využití PDT jako pomocného zdroje při zodpovídání určitých druhů konkrétních dotazů v jazykové poradně.

Harmonogram



T2-3 o----------/
T2-4 o----------/

T3
T3-1 o----------/
T3-2 o----------/
T3-3 o----------/

T4
T4-1
T4-2 o----------/
T4-3 o----------/
T4-4 o----------/
T4-5 o----------/
T4-6 o----------/

T5
T5-1 o----------/
T5-2 o----------/
T5-3 o----------/

T6
T6-1 o----------/
T6-2 o----------/
T6-3 o----------/
T6-4 o----------/
T6-5 o----------/

T7
T7-1 o----------/
T7-2 o----------/

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1,110 00, IČO: 00216208

Kalkulace předpokládaných celkových finančních výdajů projektu v r. 2004, vč. podrobné specifikace, jejich členění a zdůvodnění jednotlivých položek [3]

1. *Rozpis celkových výdajů v daném roce, tj. za všechny účastníky a jednoznačná specifikace položek hrazených z účelové dotace*

Prostředky	**Centrum** vklady zakládajících organizací + účelové prostředky požadované na zadavateli	účelové prostředky požadované na zadavateli			
		Centrum	**MFF UK**	**ZČU**	**ÚJČ**
Investice	**3 500**	**3 500**	**2 900**	**400**	**200**
Neinvestice	**31 943**	**22 681**	**19 065**	**1 839**	**1 777**
Mzdy	11 576	11 318	9 983	603	732
z toho OON	600	600	600	0	0
Režie	7 909	1 894	1 514	206	174
Ostatní	12 458	9 469	7 568	1 030	871
Celkem	**35 443**	**26 181**	**21 965**	**2 239**	**1 977**
Ostatní podrobně					
Odpisy	2 839	0	0	0	0
Cestovné a pobyty	3 048	3 048	2 638	234	176
Pojištění	4 051	3 961	3 494	211	256
DHM a NHM	843	843	586	111	146
Materiál	539	539	457	35	47
Služby	249	249	74	146	29
Další	889	829	319	293	217

Poznámka:
Tabulka odpovídá kalkulaci předpoklá. Byla opravena výpočetní chyba (rozdíl 1 tis.), výsledné částky zůstávají stejné.

2. *Specifikace a zdůvodnění výdajových položek ve vztahu k projektu*

MFF UK

Investice

Investiční prostředky v celkové výši 2 900 tis. budou v roce 2004 využity na inovaci výpočetní techniky.

[3] Tato specifikace musí obsahovat podrobný rozpis (kalkulaci) a specifikaci všech finančních nákladů/výdajů projektu a celkové částky musí odpovídat hodnotám uváděným v "excelovské" tabulce (F1E), jejich zdůvodnění musí být ve vztahu k dílčím cílům projektu (komentář ke kalkulaci).

Neinvestice

Neinvestiční prostředky pro rok 2004 budou využívány v souladu s návrhem Centra pro jeho běžný provoz. Oproti plánu nenastanou žádné výrazné změny. Z prostředků na cestovné budou hrazeny cesty na hlavní zahraniční či tuzemské konference a konference s přijatými referáty. Neinvestičními prostředky budou rovněž kryty náklady na činnost Centra, která je odrobně specifikovaná v bodech T1 až T6.

ZČU

Investice - nákup investic se uskuteční v celkové výši 400 tis. Kč.

Inovace výpočetní techniky - PC pracovních stanic a příslušenství 400 tis. Kč

Neinvestice – plán 1 840 tis. Kč

mzdy	603 tis. Kč
z toho OON	0 tis. Kč
režie	206 tis. Kč
ostatní	1 031 tis. Kč

Ostatní podrobně:

cestovné	234 tis. Kč	(konference ICASSP2004, TSD2004, ICSLP2004 ap.)
pojištění	211 tis. Kč	
DHM a NHM	111 tis. Kč	(soft pro zpracování řeči a jazyka, literatura, neinvestiční počítače pro anotační práce, přenosné disky ap.)
materiál	35 tis. Kč	(tonery, papír, kancelářské potřeby
služby	146 tis. Kč	(vložné na konference, externí služby ap.)
další	294 tis. Kč	

Struktura „další“ (celkově 294 tis. Kč) :

100 tis. Kč místnosti (nájemné, energie, teplo, telefony)
140 tis. Kč stipendia pro doktorandy a studenty
20 tis. Kč tiskové služby
34 tis. Kč opravy a údržba

Spoluúčast ZČU bude z administrativních důvodů realizována v rubrice *Režie*. Zde se v r. 2004 předpokládá vklad ZČU (na účet projektu) částky cca 251 tis. Kč. Tato částka bude využita na částečné pokrytí režijních nákladů Centra. ZČU bude v r. 2004 hradit i odpisy za nákup investičního majetku zakoupeného. Tyto odpisy však z administrativních důvodů nelze vykazovat prostřednictvím zvláštního účtu zřízeného pro spoluúčast ZČU na chodu Centra.

ÚJČ

Investice

Nákup investic se uskuteční v celkové výši 200 tis. Kč.

Investiční prostředky budou použity k inovaci výpočetní techniky - PC pracovních stanic a příslušenství

Neinvestice

Mzdy budou využity na plat jedné pracovnice jazykové poradny a na odměny třem až čtyřem studentkám zaměstnaným na částečný úvazek + odměny pro 2 pracovníky jazykové poradny.

Z položky DHM a NHM budou placeny zejména knihy, software, drobná výbava (cca 73 tis.), nábytek (cca 30 tis.) a materiál (cca 30 tis.) Cestovné bude použito na studijní pobyty a tuzemské i zahraniční konference

3. Specifikace položek hrazených z prostředků příjemce, příp. spolupříjemce

Prostředky nositele	**Celkem**	**MFF UK**	**ZČU**	**ÚJČ**
Investice	**0**	**0**	**0**	**0**
Neinvestice	**9 262**	**7 840**	**490**	**932**
Mzdy	258	172	0	86
Režie	6 015	5 113	221	681
Odpisy	2 839	2 475	239	125
Místnosti	20	0	20	0
Pojištění	90	60	0	30
Další	40	20	10	10
Celkem	**9 262**	**7 840**	**490**	**932**

V roce 2000 byly investice hrazeny (nakupovány) na všech pracovištích CKL, a na jednotlivých pracovištích jsou také odpisovány. V roce 2001 byly veškeré investice pořizovány na pracovišti nositele, tj. MFF UK (a na pracoviště spolupříjemců byly zapůjčeny), proto jsou investice z r. 2001 odpisovány pouze na MFF (na MFF byly tedy odpisy zvýšeny a na spoluřešitelských pracovištích odpovídajícím způsobem sníženy). V roce 2002 a 2003 opět jednotlivá pracoviště realizovala investice samostatně, a také je sama odepisují. Se stejným schématem – pracoviště budou nakupovat investice samostatně - se počítá i v roce 2004.
Tabulka pro rok 2004 byla oproti původnímu plánu upravena tak, aby zachycovala popsané změny v odpisech.

Poznámka:
Uvedené tabulky se shodují s položkami a částkami uvedenými v tabulce ve formátu *. xls.

Název projektu :	*Centrum komputační lingvistiky*
Řešitel:	Prof. PhDr. Eva Hajičová, DrSc.
Příjemce dotace:	Univerzita Karlova v Praze, Ovocný trh 3/5, Praha 1,110 00, IČO: 00216208

Tisková zpráva [4] ***(musí obsahovat dosažené cíle, resp. výsledky projektu – určeno pro závěrečnou zprávu do CEP):***
(max. rozsah 254 znaků)

V dne:

řešitel projektu (podpis)

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

[4] 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 1,110 00, IČO: 00216208

Příloha 1:

Seznam dat a souvisejících nástrojů získaných v rámci realizace projektu

Data získaná v rámci jednotlivých projektů Centra

- **Pražský závislostní korpus (PDT)**
  Pražský závislostní korpus (PDT), který je stěžejním výsledkem práce Centra, bude zveřejněn jako CD-ROM Pražský závislostní koprus, verze 2.0. Oproti verzi 1.0 (2001) bude obsahovat navíc 55 tisíc vět anotovaných velmi podrobně na tzv. tektogramatické rovině, a řadu nových a vylepšených nástrojů na zpracování češtiny. S finální publikací (opět prostřednictvím Linguistic Data Consortium, Univ. of Pennsylvannia, PA, USA) se počítá v závěru roku 2004.

- **Prague Arabic Dependency Treebank**
  Závislostní korpus moderní standardní arabštiny vzniká s využitím bohatých zkušeností a nástrojů získaných při vytváření PDT ve spolupráce 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.

- **VALLEX 1.0**
  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 pro nekomerční účely volně k využití, http://ckl.mff.cuni.cz/zabokrtsky/vallex/1.0/.

- **Český anotovaný korpus**
  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. V prvních měsících roku 2004 bude dokončena konverze vnitřního kódování a morfologického anotování; první verze korpusu bude následně zveřejněna – dále viz http://ckl.mff.cuni.cz/~sgd/CAC.html.

- **Prague Czech-English Dependency Treebank**
  Prague Czech-English Dependency Treebank (PCEDT) je paralelní, česko-anglický závislostní korpus, který bude v průběhu roku 2004 vydán v Linguistic Data Consortium (LDC). 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.

- **Korpusy spontánních promluv projektu MALLACH (ZČU Plzeň)**
  - **Český korpus anotovaných výpovědí lidí přeživších holocaust:**
    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:**
  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):**
  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
  (Korpusy projektu MALACH jsou veřejně nedostupné!)

- **Czech Broadcast News Corpus**
  (vyjde v LDC v lednu 2004, všechny podklady odevzdány v nakladatelství)
  řečový signál: 22,05 kHz, 16 bitů
  rozsah korpusu: cca 50 hod vysílání
  stanice: ČRo1, ČRo2, ČRo3, ČTV, Prima
  (Tento korpus se bude prodávat v nakladatelství LDC v r. 2004)

- **Old-Church Slavonic Corpus (OCS)**
  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. to cca 17. století. Přístup k datům bude umožněn přes http://ckl.ms.mff.cuni.cz/~ribarov.

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. Používá se v procesu značkování PDT (popis viz TR-2001-12, 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.
Netgraph je pro akademické účely volně k dispozici na internetu, včetně podrobné dokumentace – viz http://quest.ms.mff.cuni.cz/netgraph.
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 (viz zahraniční cesty).

- **Syntaktické analyzátory češtiny** ("parsery")
  V CKL se paralelně vyvíjí 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. Bude volně šiřitelný pro nekomerční účely. V blízké době chystáme nejen vyvěšení balíčku s parserem ke stažení na našich webových stránkách, ale také pokusné zprovoznění on-line analýzy 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 budou podrobně popsány v dokumentaci k Prague Czech-English Dependency Treebank, který bude vydán v následujícím roce v LDC (viz výš).

- **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 nespisovné tvary slov 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 pomocí studentů vyvinut programový balík ACT (Annotated Corpora of Text) - jazykově nezávislí nástroj pro vytváření anotovaných korpusů s řadou speciálních funkcí pro zachycení jazykových víceznačností a variant. V rámci ACT je možné lemmatizovat, dezambiguovat (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. e 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.

Příloha 2:

Seznam publikací

1. 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* (in press). Supported by LN00A063 and MSM113200006.

2. Bojar, Ondřej (2003): **Building Subcorpora Suitable for Extraction of Lexico-Syntactic Information**. In *Proceedings of the Student Session, ESSLLI*, pp. 25--34. Supported by LN00A063 and MSM113200006.

3. Bojar, Ondřej (2003): **Ax - Systém pro automatizovanou extrakci lexikálně-syntaktických údajů**. In *Malý informatický seminář* (to appear). Supported by LN00A063 and MSM113200006.

4. Bojar, Ondřej (2003): **Towards Automatic Extraction of Verb Frames**. In *Prague Bulletin of Mathematical Linguistics*, pp. 101--120. MFF UK. Supported by 300/2002/A INF-MFF and GA201/02/1456.

5. Bojar, Ondřej; Brom, Cyril; Hladík, Milan; Vejlupek, Mikuláš; Toman, Vojtěch; Voňka, David (2003): **Simulátor přirozeného prostředí lidského světa**. In *Malý informatický seminář* (to appear). Supported by LN00A063 and MSM113200006.

6. Böhmová, Alena; Hajičová, Eva (2003): **Large language data and the degrees of automation**. In *CIL 17 Proceedings* Benjamins (in press). Supported by LN00A063 and MSM113200006.

7. Camuglia, Giuseppe; Camuglia Ribarov, Monia; Ribarov, Kiril (2003): **Computer Processing of a Clopen Language System: Old-Church Slavonic**. In *Linguistica Computazionale* Istituti Editoriali e Poligrafici Internazionali. Supported by VS 96151 and GA405/96/K214.

8. 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"* Supported by LN00A063 and MSM113200006.

9. Cinková, Silvie (2003): **Belegsuche bei der lexikographischen Bearbeitung von selten gebrauchtem Wortschatz**. In *Das Wort. Germanistisches Jahrbuch 2003*, pp. 353--365. Deutscher akademischer Austauschdienst.

10. 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. Supported by LN00A063 and MSM113200006.

11. 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. Supported by ME642 and NSF IIS-0121285 and LN00A063.

12. 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* Supported by LN00A063 and MSM113200006.

13. Gramatovici, Radu (2003): **On the Recognizing Power of Non-Expansive Go-Through Automata**. In *Annals of the Bucharest University* Supported by LN00A063.

14. Hajič, Jan; Homola, Petr; Kuboň, Vladislav (2003): **A Simple Multilingual Machine Translation System**. In *Proceedings of Machine Translation Summit IX*, pp. 157--164. Supported by LN00A063 and MSM113200006.

15. 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. Supported by GA405/03/0913 and LN00A063.

16. Hajič, Jan; Kuboň, Vladislav (2003): **Tagging as a Key to Successful MT**. In *MIS 2003* Supported by LN00A063 and MSM113200006.

17. 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. Supported by GA405/03/0913 and LN00A063.

18. 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* (submitted). Supported by LN00A063.

19. 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. Supported by GA405/03/0913 and LN00A063.

20. Hajičová, Eva (2003): **Information structure and syntactic complexity**. In *Investigations into formal Slavic linguistics*, pp. 169--180. Peter Lang. Supported by LN00A063 and MSM113200006.

21. 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. Supported by LN00A063 and MSM113200006.

22. Hajičová, Eva (2003): **Contextual boundness and discourse patterns**. In *CIL 17 Proceedings* Benjamins (in press). Supported by LN00A063 and MSM113200006.

23. 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. Supported by LN00A063 and MSM113200006.

24. 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". Supported by LN00A063 and MSM113200006.

25. Hajičová, Eva; Sgall, Petr (2003): **Information Structure, Translation and Discourse**. In *Textologie und Translation*, pp. 107--123. Gunter Narr. Supported by LN00A063 and GA405/96/K214.

26. 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. Supported by LN00A063 and GA405/96/K214.

27. 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. Supported by LN00A063 and MSM113200006.

28. Hlaváč, Václav; Hlaváčová, Jaroslava (2003): **Rozpoznávání jako jeden z přístupů porozumění složitým jevům**. In *Softwarové noviny*

29. 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. Supported by LN00A063 and MSM113200006.

30. 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. Supported by LN00A063 and MSM113200006.

31. Holub, Martin; Straňák, Pavel (2003): **Approaches to Building Semantic Lexicons**. In *WDS'03 Proceedings of Contributed Papers, Part I*, pp. 173--178. MATFYZPRESS. Supported by LN00A063 and MSM113200006.

32. 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* (in press). Supported by LN00A063 and MSM113200006.

33. Honetschläger, Václav (2003): **Using a Czech Valency Lexicon for Annotation Support**. In *Proceedings of Text, Speech and Dialogue 2003*, pp. 120--126. Springer. Supported by GA405/03/0913 and LN00A063 and MSM113200006.

34. 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. Supported by LN00A063.

35. Klímová, Jana; Kolářová-Řezníčková, Veronika (2003): **Využití ČNK a PZK pro ověřování valenčních vlastností deverbativních substantiv se zabudovanou rolí**. In *Slovanské jazyky v počítačovom spracovaní* Supported by GA405/03/0913 and GA405/03/0377.

36. Kocanda, Jiří (2003): **Statistical Parsing**. In *WDS'03 Proceedings of Contributed Papers, Part I*, pp. 161--166. MATFYZPRESS. Supported by LN00A063 and MSM113200006.

37. 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. Supported by GA405/03/0913 and LN00A063.

38. 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. Supported by LN00A063 and MSM113200006.

39. Kučová, Lucie; Kolářová, Veronika; Žabokrtský, Zdeněk; Pajas, Petr; Čulo, Oliver (2003): **Anotování koreference v Pražském závislostním korpusu**. MFF UK. Supported by Německý akademický výměnný program and LN00A063.

40. 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. Supported by LN00A063 and MSM113200006.

41. Květoň, Pavel (2003): **Language for Grammatical Rules**. MFF UK. Supported by GA405/03/0913 and LN00A063 and MSM113200006.

42. Lopatková, Markéta (2003): **Valency in the Prague Dependency Treebank: Building the Valency Lexicon**. In *Prague Bulletin of Mathematical Linguistics*, pp. 37--60. MFF UK. Supported by LN00A063.

43. Lopatková, Markéta (2003): **O homonymii předložkových skupin v češtine (Co umí počítač?)** Karolinum. Supported by LN00A063 and MSM113200006.

44. Lopatková, Markéta; Panevová, Jarmila (2003): **Valence vybraných skupin sloves (k některým slovesům dandi a recipiendi)**. In *Čeština -- univerzália a specifika*, pp. x01--x10. Lidové noviny (in press). Supported by LN00A063 and MSM113200006.

45. Lopatková, Markéta; Žabokrtský, Zdeněk (2003): **Testování konzistence a úplnosti valenčního slovníku českých sloves**. In *Proceedings of ITAT 2003* (in press). Supported by LN00A063 and MSM113200006.

46. Lopatková, Markéta; Žabokrtský, Zdeněk; Skwarska, Karolina; Benešová, Václava (2003): **VALLEX 1.0 Valency Lexicon of Czech Verbs**. MFF UK. Supported by LN00A063.

47. 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. Supported by GA405/03/0913 and LN00A063 and MSM113200006.

48. 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. Supported by LN00A063 and MSM113200006.

49. Panevová, Jarmila (2003): **O jednom typu kauzativní konstrukce v češtině**. In *Etudes linguistiques Romano-Slaves offertes a Slanislaw Karolak*, pp. 379--385. Oficyna Wydavnicza "Edukacja". Supported by MSM113200006.

50. Plátek, Martin; Lopatková, Markéta; Oliva, Karel (2003): **Restarting Automata: Motivations and Applications**. In *Proceedings of the workshop "Petrinetze"*, pp. 90--96. Technische Universitaet Muenchen. Supported by LN00A063 and MSM113200006.

51. 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. Supported by LN00A063 and MSM113200006.

52. 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. Supported by NSF IIS-0122466 and MSM234200004 and LN00A063.

53. 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. Supported by LN00A063 and MSM113200006.

54. 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. Supported by ME642 and LN00A063.

55. Ribarov, Kiril; Camuglia, Monia (2003): **Incorporation of Old-Church Slavonic Card-Files into a Corpus**. In *Scripta & e-Scripta* Institute of Literature, Bulgarian Academy of Sciences. Supported by LN00A063 and MSM113200006.

56. 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. Supported by LN00A063 and MSM113200006.

57. Semecký, Jiří (2003): **Semantic Word Classes Extracted from Text Clusters**. In *WDS'03 Proceedings of Contributed Papers, Part I*, pp. 167--172. MATFYZPRESS. Supported by LN00A063 and MSM113200006.

58. 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. Supported by LN00A063 and MSM113200006.

59. Sgall, Petr (2003): **Slavistics and the history of topic-focus studies**. In *Investigations into formal Slavic linguistics*, pp. 201--212. Peter Lang. Supported by LN00A063 and MSM113200006.

60. Sgall, Petr (2003): **Lingvistické ohlédnutí za dvacátým stoletím**. In *Český jazyk a literatura*, pp. 157--164. SPN & Fortuna. Supported by LN00A063 and MSM113200006.

61. Sgall, Petr (2003): **Dynamics in the meaning of the sentence and of discourse**. In *Meaning: The Dynamic Turn*, pp. 169--184. Elsevier Science Ltd. (in press). Supported by LN00A063.

62. 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". Supported by LN00A063 and MSM113200006.

63. 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. Supported by LN00A063 and MSM113200006.

64. Sgall, Petr (2003): **Introductory remarks to the Workshop on Discourse Structures at the 17th International Congress of Linguists, Prague**. In *CIL 17 Proceedings* Benjamins (in press). Supported by LN00A063 and MSM113200006.

65. Sgall, Petr (2003): **Types of Languages and the Simple Pattern of their Core**. In *CIL 17 Proceedings* Benjamins (in press). Supported by LN00A063 and MSM113200006.

66. Veselá, Kateřina; Havelka, Jiří (2003): **Anotování aktuálního členění věty v Pražském závislostním korpusu**. MFF UK. Supported by LN00A063.

67. 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. Supported by LN00A063.

68. Veselá, Kateřina; Peterek, Nino; Hajičová, Eva (2003): **Some observations on contrastive topic in Czech spontaneous speech**. In *CIL 17 Proceedings* Benjamins (in press). Supported by LN00A063 and MSM113200006.

69. 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. Supported by LN00A063.

70. Žabokrtský, Zdeněk; Smrž, Otakar (2003): **Arabic Syntactic Trees: from Constituency to Dependency**. In *EACL 2003 Conference Companion*, pp. 183--186. Association for Computational Linguistics. Supported by LN00A063 and MSM113200006.