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

Průběžná zpráva o realizaci projektu[1]

Povinná osnova zprávy:

1. Stručný přehled dílčích cílů projektu splněných v uplynulém období

V prvních šesti měsících realizace projektu byly splněny následující cíle původního projektu:

(A) Teoretická komputační lingvistika:

Byly probrány některé otevřené otázky popisu syntaktické stavby češtiny, především tzv. valenčních rámců sloves a podstatných jmen.

(B) Uplatnění velkých souborů dat pro jazykovou analýzu:

1. Využití korpusu pro Czech Language Help:
   Byly stanoveny základní zásady pro využití korpusu pro tento účel.
2. Vytvoření specializované jazykové databáze:
   Byl vytvořen první náčrt celkové struktury a předložen k diskusi spoluřešitelům.
3. Práce se soustředila na upřesňování instrukcí pro anotátory pro přiřazování valenčních rámců deverbativ.
4. Materiálové sondy do Pražského závislostního stromu, týkající se jevů, které jsou buď v dlouhodobému vývojovém pohybu v češtině, nebo aktuální záležitostí jazykové poradny, a spolu s tím ověřování správnosti anotování korpusu (pokračování prací započatých v r. 2000). Univerzália a specifika internetové jazykové poradny jako zvláštního registru/registrů elektronické komunikace.
5. Výzkum v oblasti syntaxe a sémantiky češtiny s využitím Pražského závislostního stromu, verifikace anotačních procedur.
6. Byly modifikovány instrukce pro anotátory pro přiřazování hodnot atributu 'aktuálního členění' a tento návrh byl prezentován v přednášce na mezinárodní konferenci COLING 2000.
7. Byly prostudovány a v přednášce na mezinárodní konferenci Text, Speech and Dialogue 2000 prezentovány podmínky vypouštění členů při přechodu z hloubkové na povrchovou strukturu věty.
8. Byl navržen a v přednášce na mezinárodní konferenci COLING 2000 prezentován návrh na specifikaci hodnot atributu vyhrazenému koreferenčních vztahů.

(C) Metodologie a kombinace různých přístupů k jazykovému modelování:

9. Byl navržen první, tentativní model kombinace strukturních (pravidlových) a statistických metod pro morfologické značkování.
10. Byl dokončena první česká monografie o počítačové morfologii a morfologické desambiguaci.
11. Byly studovány nekontrolované metody morfologického a lexikálního popisu; toto studium bude pokračovat v následujícím roce projektu.

[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).

12. Byl experimentálně implementován systém automatického překladu z češtiny do slovenštiny jako ověření možnosti efektivního robustního překladu mezi typologicky velmi blízkými jazyky.

(D) Matematické a komputační základy:

Byly evaluovány výsledky několika prvních etap značkování na rovině podkladové (hloubkové) syntaxe a výsledek byl prezentován v přednášce na mezinárodní konferenci Text, Speech and Dialogue 2000.

(E) Rozpoznávání řeči

1. Jazykové modelování češtiny, dialogové systémy: Byl natrénován akustický model s využitím nového korpusu mluvené řeči (Czech Broadcast Speech Corpus). Zmíněný korpus obsahuje zprávy vysílané českými rozhlasovými a televizními stanicemi v období únor až květen 2000, které byly zpracovány na ZČU a poskytnuty CKL k dispozici. Před vlastním trénováním byl korpus rozdělen na 2 části – pouze první část (cca 21 hodin transkribované řeči) byla použita pro trénování, druhá část (cca 5 hodin) byla odložena stranou pro pozdější testování, tj. určení WER.
   Pomocí speciálních nástrojů byly jednotlivé komponenty systému pro rozpoznávání řeči (akustický model, běžný back-off bigramový jazykový model a výslovnostní slovník) převedeny do formátu konečných automatů a pomocí dekodéru byla pro každou testovací promluvu vygenerována takzvaná lattice, což je vlastně orientovaný graf, který obsahuje slova považovaná dekodérem za nejpravděpodobnější. Z těchto lattice byl extrahován seznam *N*-nejpravděpodobnějších posloupností slov (*N*-best list), který může být později reskórován pomocí nových jazykových modelů, které budou pro češtinu vhodnější než standardní bigramový model.
   Kromě výše uvedených prací byla testována metoda konstrukce jazykových modelů, která se snaží vyřešit nedostatek trénovacích dat (a z toho plynoucí nedostatečné natrénování parametrů modelu) shlukováním slov do tříd. Pro rozdělení slov do tříd byly použity dvě různé metody - první z nich odvozuje jednotlivé třídy přímo z trénovacích dat, zatímco druhá využívá informaci o slovním druhu daného slova.
2. Význam větné prozodie ve vztahu k AČ a příprava prosodického korpusu. Byl vyhledán vhodný soubor prosodicky charakteristických vět. Tyto věty budou namluveny více mluvčími a opatřeny značkami jak na úrovni akustické (průběh základního tónu, trvání, energie), tak i na úrovni lingvistické (slabiky, morfologie, větný důraz). Soubor bude dále rozširován o spontánní promluvy a dialogy, na kterých budou ověřovány závislosti zjištené ve čtených datech korpusu.
3. Jazykové modelování (využití modelů s lineární historií pro flexivní jazyk): Pro jazykové modelováni jsou používány různé N-gramové modely, převážně založené na dekompozici slovních tvarů na kmen, koncovka a morfologická značka, lemma. Pro model založený na morfologických značkách bude upraveno morfologické značkování tak, aby se při běhu používal pouze levý kontext, což umožní zakomponovat celý model do výsledného "Real Time Speech Recogniser".

(F) Využití vícejazyčných zdrojů:

Byl zahájen výzkum založený na využití paralelních korpusů pro formulaci a realizaci experimentálního systému strojově podporovaného překladu.

2. Personální a organizační zabezpečení činnosti Centra (aktuální stav)

K 1.7.2000 bylo založeno Centrum komputační lingvistiky MFF UK; bylo ustaveno jeho vedení (vedoucí prof. PhDr Eva Hajičová, DrSc, zástupce vedoucí mgr. Barbora Hladká-Vidová, PhD) a jeho vnitřní organizační struktura. CKL se člení na čtyři oddělení:

(a) teoretické otázky jazykové analýzy a budování Pražského závislostního korpusu (prof. PhDr Jarmila Panevová, DrSc),

(b) zpracování mluvené řeči (prof. ing. Josef Psutka, CSc),

(c) jazykové modely, především stochastické (RNDr Jan Hajič, PhD),

(d) aplikační systémy (RNDr Vladislav Kuboň).

Bylo uskutečněno několik pracovních schůzek vedení CKL, na nichž byly projednány otázky odborné práce i administrativních pravidel spolupráce.

Složení pracovního týmu z hlediska kvalifikace ve vztahu k pracovní náplni v Centru je velmi vyvážené a ve shodě s původním návrhem projektu.

Na řešení projektu se podílí:

- 18 pracovníků s úvazkem rovným nebo vyšším než 0,7, a
- 20 pracovníků s úvazkem nižším než 0,7.

Rovněž z hlediska věku pracovníků je 32 pracovníků mladších než 35 let (jejich kapacitní podíl na projektu je 84%)

3. Přístrojové vybavení a technické zabezpečení činnosti Centra

V zahajovacím roce jsme Centrum technicky zabezpečili následujícím vybavením:

- 8 pracovních stanic P III (některé dvouprocesorové), 7 z nich s monitory,
- 1 datový server RAID 5-Xeon, 500 GB,
- 1 notebook,
- 2 osobní počítače,
- byl proveden upgrade tří počítačů
- 2 switchované karty (100 Mbps porty),
- 4 gigabitové karty pro připojení Centra na páteř budovy MFF UK,
- 1 karta do páteřního racku,
- 1 switch do místnosti centra,
- 1 switch pro centrum na ZČU
- výstupní a prezentační zařízení: 1 barevná laserová oboustranná tiskárna, 1 prezentační promítačka Philips.

Nákup pracovních stanic byl realizován z důvodů potřeby permanentního a časově velmi náročného trénování akustických a jazykových modelů. Též časově i výpočetně mimořádně náročné experimenty s rozpoznáváním řeči zcela využívají výkon dvou-procesorových pracovních stanic. Jednoprocesorové pracovní stanice s nižším výkonem (a někde i se sdíleným monitorem) jsou využívány pro přípravu dat, resp. programů. Switch-e/100MHz byly zakoupeny podle plánu z důvodů konsolidace počítačové sítě a připojení Centra na rychlou 100MHz síť.

Činnost Centra je ze 70-80% zaměřena na experimenty s jazykovým materiálem s použitím metod stochastického modelování typicky s milióny až miliardami parametrů (připomeňme, že např. pro spektrální modely ve fyzice nebo chemii, považované za složité, se používá ne více než několik desítek až stovek parametrů - křivka normálního rozdělení má obvykle pouhé dva). Pro takové experimenty je třeba mít k dispozici silnou počítačovou lokální síť s mimořádně velkým

"skladovacím prostorem" pro prvotní data (texty a zvukové nahrávky), zpracovávaná data a mezivýsledky (v průběhu zpracování nároky na diskový prostor ještě řádově rostou) a do jisté míry i pro výsledky.

Stávající kapacity jsou plně využívány. Síťové a serverové prvky byly, kromě splnění i výše uvedených úkolů, vybírány tak, aby rentabilně umožnily i další rozšíření současného počítačového parku v souladu s návrhem Centra.

4. Spolupráce Centra

Centrum od samého počátku své činnosti úzce spolupracuje především s Ústavem Českého národního korpusu FF UK, jehož soubory jsou zdrojem dat pro výzkum CKL. Tato spolupráce navazuje logicky na činnost Laboratoře počítačového zpracování jazykových dat při ÚFALu MFF UK, zřízené v rámci projektu MŠMT v r. 1996; tato Laboratoř končí podle plánu svou činnost koncem r. 2000 a šestiměsíční souběh prací na obou pracovištích umožnil bezproblémovou návaznost výzkumu, a to jak z hlediska obsahového, tak i z hlediska personálního. Zájem o spolupráci s CKL mají i další česká pracoviště obdobného charakteru, byť se specifickým zaměřením na ten či onen aspekt výzkumu, jako je např. Ústav teoretické a komputační lingvistiky FF UK, pracoviště komputační lingvistiky Fakulty informatiky Masarykovy univerzity v Brně, pracoviště akustické analýzy na univerzitě v Liberci, atd.

V mezinárodním měřítku se realizuje především velmi intenzívní spolupráce s pracovišti Johns Hopkins University (prof. F. Jelinek z této univerzity přijede na jednosemestrální, možná i dvousemestrální přednáškový pobyt do Prahy jako hostující profesor CKL, v současné době je jako hostující profesor na JHU jeden z pracovníků CKL, dr. Jan Hajič, a jako 'research fellow' zástupkyně vedoucí CKL mgr. Barbora Hladká, PhD).

Ve spolupráci s aplikační sférou se rozvíjí velmi slibně spolupráce s firmou TRADOS v oblasti strojově podporovaného překladu.

5. ***Doktorské studijní programy***

CKL se významnou měrou podílí na doktorském programu I3 - Informatika - Počítačová lingvistika na Matematicko-fyzikální fakultě UK (školiteli nebo konzultanty jsou pracovníci CKL prof. Eva Hajičová, prof. Jarmila Panevová, prof. Petr Sgall, a dr. Jan Hajič) a na doktorském programu v oboru Kybernetika na Fakultě aplikovaných věd, Západočeská univerzita (prof. Josef Psutka).

V celkovém počtu se uvedené osoby angažují ve vedení 28 doktorandů.

6. Podpora mladých výzkumných pracovníků

Podpora mladých výzkumných pracovníků je jednou z priorit programu CKL. Projevuje se především podporou všestranných odborných kontaktů zejména se zahraničními vědci, ať již na konferencích u nás nebo na studijních či pracovních pobytech (srov. výše bod 4 o zahraniční spolupráci).

(a) Účast na mezinárodních konferencích:

Mezinárodní kongres počítačové lingvistiky COLING 2000, Saarbrücken, Německo, 29.7.-7.8.

- Dan Zeman: přednesení vlastního referátu na konferenci v Saarbrückenu i na semináři v Lucemburku
- Alena Böhmová: přednesení vlastního referátu na semináři v Lucemburku
- Markéta Ceplová: spoluautorka přednesného referátu na konferenci

Mezinárodní výroční konference Societas linguistica Europaea, Poznaň, Polsko, 30.8.- 2.9.

- Markéta Straňáková: přednesení vlastního referátu

Mezinárodní konference Text, Speech and Dialogue 2000, Brno, 13.9.-15.9.

- Markéta Straňáková: přednesení vlastního referátu
- Nino Peterek: autor posteru
- Dan Zeman: autor posteru
- Jan Cuřín, Pavel Krbec

Mezinárodní konference o komputační lexikologii, COMLEX, Řecko, 20.9.-25.9.

- Roman Ondruška

Mezinárodní výroční konference světové společnosti Association for Computational Linguistics, Hong Kong, Čína, 1.10.-10.10.

- Alena Böhmová: přednesení vlastního referátu
- Jiří Havelka: účast na semináři Research Advances in Natural Language Processing and Information Retrieval

(b) Účast na letní škole:

European Summer School in Logic, Language and Information, Birmingham, Velká Británie, 7.8. - 17.8.

- Veronika Řezníčková, Jiří Havelka

(c) Univerzita Birmingham, Lingvistická korpusová skupina (prof. Teubert)

- studijní cesta dvou pracovnic ÚJČ

Poznámka: Většina cest byla částečně hrazena též ze zdrojů pořádající organizace nebo z grantu GAČR a MŠMT.

7. Způsob zpřístupnění výsledků a výstupů centra veřejnosti

(a) Byl uspořádán seminář s mezinárodní účasti na zahájení činnosti Centra, na němž zahraniční účastníci přednesli přednášky o svých projektech a představách automatického zpracování přirozeného jazyka, a pracovníci CKL pak představili východiska svého projektu a jeho směřování.

(b) Ve spolupráci s řešiteli projektu MŠMT "Laboratoř počítačového zpracování jazykových dat" a komplexního projektu GAČR KP-214 bylo připraveno vydání kompaktního disku o Pražském závislostním korpusu obsahujícího jak specifikaci značkování korpusu, tak i data a programové nástroje vyvinuté především v uvedených dvou projektech.

(c) Byla zkompletována a doplněna webová stránka Pražského závislostního korpusu (v anglickém znění) http://ckl.mff.cuni.cz .

(d) Výzkum v Centru byl také prezentován na řadě mezinárodních konferencí a na přednáškových pobytech; pokud jde o mladé pracovníky, uvádíme soupis těchto pobytů v bodě 6 výše, dále dodáváme:

Mezinárodní kongres počítačové lingvistiky COLING 2000, Saarbrücken, Německo, 29.7.-7.8.

- Petr Sgall: člen panelu; účast na zasedání jako člen mezinárodního výboru počítačové lingvistiky; přednesení vlastního referátu
- Eva Hajičová: přednesení vlastního referátu

Mezinárodní konference o kontrastivním studiu sémantiky, Cambridge, Velká Británie, 9.9.-13.9.

- Eva Hajičová: přednesení pozvaného plenárního referátu

Mezinárodní konference Text, Speech and Dialogue 2000, Brno, 13.9.-15.9.

- Petr Sgall: člen programového výboru

Mezinárodní výroční konference světové společnosti Association for Computational Linguistics, Hong Kong, Čína, 1.10.-10.10.

- Eva Hajičová: účast na zasedání jako členka hlavního výboru společnosti

Mezinárodní konference ICSLP´2000, Peking

- Josef Psutka, účast na konferenci a přednesení referátu

Zasedání Komise pro gramatickou stavbu slovanských jazyků při Mezinárodním komitétu slavistů, Zurich, Švýcarsko, 21.10.-24.10.

- Jarmila Panevová: členka Komise

Fakulta informatiky, Masarykova univerzita Brno

- Eva Hajičová: 17.10., pozvaná přednáška na celofakultním informatickém semináři o poznatcích z nejnovějšího vývoje počítačové lingvistiky

Název projektu : *Centrum komputační lingvistiky*
Řešitel: prof. PhDr. Eva Hajičová, DrSc.
Nositel: 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. 2001[2]

Práce budou probíhat podle harmonogramu stanoveného v původním návrhu projektu, a to v šesti větvích:

(A) V oblasti teoretických aspektů komputační lingvistiky půjde o podrobnější studium vztahů věty k nadvětnému kontextu, především z hlediska aktuálního členění větného. Zvláštní pozornost bude věnována v této souvislosti i otázkám českého slovosledu.

(B) (i) V návaznosti na práce provedené v prvním roce projektu bude pokračovat značkování Pražského závislostního korpusu na úrovni tektogramatické, a to v obou proudech, velkém i vzorovém, s cílem dosáhnout koncem roku objemu 10 tisíc označkovaných vět ve velkém souboru a alespoň tisíc vět v souboru vzorovém.

(ii) Budou pokračovat práce na specializované jazykové databázi, a to především z hlediska vytyčení její celkové struktury, databáze a kompilace a evaluace datových zdrojů.

(iii) Značkovaných souborů Pražského závislostního korpusu bude využito ke studiu jednotlivých jazykových jevů tak, aby bylo možné zpřesnit instrukce pro anotátory a zároveň dospět k novým teoretickým poznatkům v daných oblastech; půjde zejména o otázky valence deverbativ ve vztahu k valenci základních sloves, formulaci poloautomatické procedury značkování uzlů stromu vzhledem k aktuálnímu členění věty, formulaci přesnějších pravidel pro rekonstrukci uzlů a zpřesnění instrukcí pro naplňování koreferenčních atributů.

(C) V oblasti metodologie půjde o pokračování výzkumu o možnostech propojení empirických a strukturních metod disambiguace a o uplatnění prvního návrhu takového propojení na morfologické značkování korpusu. Dále bude pokračovat studium tzv. nekontrolovaných metod morfologického a lexikálního popisu a zdokonalování metody strojového překladu mezi blízkými jazyky, ověřované i experimentálně.

(D) V oblasti komputační pak půjde o soustavnou počítačovou (softwarovou) podporu prací uvedených v předchozích bodech a o vývoj příslušných algoritmů.

(E) (i) V následujícím období budou maximální síly soustředěny zejména na seznámení se prací a "obsluhou" dekodéru AT&T, jež umožní potřebné experimenty s velkými slovníky (řádově desítky tisíc slov). Předpokládáme, že budou provedeny experimenty s reskórováním *N*-best listů pomocí nových jazykových modelů navržených speciálně pro češtinu. Chceme především otestovat model založený na kmenech a koncovkách, model vyhlazený pomocí lineární interpolace a lineární interpolace s bucketingem. Pokud některý z nově navržených jazykových modelů výrazně zlepší WER, budeme se snažit jej převést rovněž do podoby konečného automatu a použít přímo jako komponentu dekodéru. Výsledky těchto experimentů bychom chtěli prezentovat na některé mezinárodní konferenci.

(ii) Budou pokračovat práce na studiu vztahu větné prozodie a aktuálního členění větného, především pak z hlediska možností charakterizovat rozdíly mezi různými typy přízvuku (intonační centrum, kontrastivní přízvuk atd.)

(F) Po počátečních krocích v prvním roce projektu bude evaluován navržený shlukový přístup ke strojovému překladu, a to ve spolupráci s industriálním partnerem a bude i nadále upřesňován přístup ke strojovému překladu mezi blízce příbuznými jazyky.

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

V roce 2001 bude Centrum organizovat tyto akce:

(i) ve spolupráci s Centrem Viléma Mathesia mezinárodní dvoutýdenní jarní školu komputační lingvistiky, na jejíž organizaci i obsahové náplni se budou podílet především mladí pracovníci Centra (březen 2001);

(ii) dvoutýdenní letní školu v rámci projektu Evropské unie ELSNET, kde bude přednášet jeden z pracovníků Centra a organizovat ji budou vesměs mladí pracovníci Centra;

(iii) budou připraveny a distribuovány tři Technické zprávy (březen, červen, září);

(iv) v říjnu bude uspořádán Den otevřených dveří, na němž bude odborná veřejnost seznámena s výsledky výzkumu Centra v roce 2001;

(v) v lednu bude uspořádán společný týdenní workshop řešitelů projektu.

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

Tisková zpráva [3]

Rozbor otevřených otázek popisu syntaktické stavby.
Návrh prvního modelu kombinace strukturních a statistických metod.
Příprava řečových a jazykových dat a vlastní experimenty s rozpoznávání mluvené češtiny.
Výzkum založený na využití paralelních korpusů pro strojový překlad.

V dne:

řešitel projektu
(podpis)

nositel projektu
(razítko a podpis statut .zást. nositele)

[3] 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!¨.