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

Průběžná zpráva o realizaci projektu[1]

V druhém roku realizace projektu byly splněny všechny cíle původního plánu – níže uvádíme podrobnosti plnění jednotlivých úkolů.

1. Stručný přehled dílčích cílů projektu splněných v uplynulém období

(A) TEORETICKÁ A KOMPUTAČNÍ LINGVISTIKA

V souvislosti s dopracováním další verze tektogramatického manuálu (červen-říjen 2001) musely být dořešeny některé teoretické a empirické problémy české syntaxe.

- vedle vytvoření koncepce elektronického valenčního slovníku českých sloves (viz níže) byla v r. 2001 věnována pozornost především následujícím okruhům:
  - **řešení "složených" časových a místních určení** (syntagmat), jde o typy *včera ráno, letos v únoru, těsně po Vánocích, dva měsíce před porodem; v Praze na Vyšehradě, dole pod podlahou, daleko od Prahy, pět minut od pláže, na sever od Alp*. V těchto konstrukcích není snadné zjistit, který člen je řídící a který závislý. Na příkladech(1) a (2) (na rozdíl od příkladů (3) a (4)) lze běžnou metodou vypustitelnosti zjistit, který člen je řídící a který závislý: vypuštěním adverbia v (2) získáme negramatickou větu. V příkladech (3) a (4) budeme proto vystupovat analogicky podle prototypu (1) a (2) a za řídící pokládat určení „dva měsíce“, ačkoli vypustitelnost jedné části zkoumaného syntagmatu je různá, neboť je podmíněna slovesem.

    (1) Toto místo leží severně (od Alp)

    (2) *Toto místo leží (severně) od Alp

  V případech (1) a (2) je třeba postupovat analogicky podle příkladu (3) a (4): kde *od Alp* se jeví jako nevypustitelné a tedy řídící.

    (3) Proležela dva měsíce (před porodem) v nemocnici

    (4) Přihodilo se jí to (dva měsíce) před porodem

  - **konstrukce se slovesem "být"** vykazují jistá specifika, zejména následuje-li vedlejší věta přísudková a doplňková. Byla navržena pravidla, jakými prostředky zachytit v tektogramatické stromové struktuře (TGTS) rozdíly mezi větami (5) a (6), jak anotovat doplňky jako(*být)rád, sám* včetně případů jejich souvýskytu

    (5) Je to vedoucí (takový), jakého jsme si přáli

    (6) Je to vedoucí, kterého jsme si přáli

  - **cizojazyčným frázím** uvnitř českého textu se přiděluje **zvláštní funktor**
  - ve vzorovém souboru byl zaveden nový **gramatém pro dispoziční modalitu** v konstrukcích jako (7). (7) a (8) se zatím liší pouze ve vzorovém souboru:

    (7) Matematika se mu studuje snadno.

[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).

(8) Matematiku studuje snadno.

- bylo provedeno detailní zkoumání **předložkových výrazů lokálních a směrových**, zahrnuty byly i sekundární předložkové výrazy. Na základě tohoto výzkumu bude upraven soubor syntaktických gramatémů u místních a směrových určení s teoretickým rozlišením nepříznakových členů pro jednotlivé významy, z nichž lze vyvodit jejich sekundární varianty
- u **koordinačních skupin** byly řešeny případy **víceznačností** a byl navržen způsob, jak v případě nejednoznačností postupovat. Zvolená notace pro anotace TGTS skýtá možnost rozlišení "společného" rozvití koordinovaných členů od rozvíjení jednotlivých členů koordinace (v konstrukci *staré stoly a židle* bude ve struktuře rozlišeno, zda adjektivum rozvíjí obě jména nebo pouze jméno nejbližší)
- bylo provedeno rozlišení, kdy formálně **anaforická zájmena** slouží jednou jako odkazovací slova, sloužící jako korelativní spojení vedlejší věty s větou řídící, a kdy jsou samostatnými členy věty s deiktickou nebo anaforickou platností. Víceznačnost je přítomna např. v (9)

(9) Mluvili o tom, kdo u nás bude bydlet

- byla zpracována podrobná pravidla, **jak anotovat přímou řeč** ve vztahu k její uvozovací části. Rozhodující přitom je typ uvozovacího slovesa a jeho valenční rámec (jeho naplněnost nebo nenaplněnost v dané větě). Pokud přímou řeč uvozuje sloveso, které nepatří ani vzdáleně mezi verba dicendi, doplňuje se přechodník s významem "řka/řkouc/řkouce", přímá řeč je pak Efektem k tomuto elidovanému tvaru slovesa pravení, srov. příklad (10)

(10) Dcera se ušklíbla: "Matka ti to jistě připraví."

Tam, kde nelze takový výraz doplnit a nejde ani o valenční člen, zachycuje se přímá řeč jako parenteze

- byla vypracována pravidla, jak zachycovat **větné členy**, které jsou **ve vztahu reciprocity**. Tato otázka je z hlediska náležitého zachycení valenční struktury slovesa podstatná. V případě reciprocity se některý z aktantů „přesouvá“ do pozice, která je jinak syntakticky hierarchizovaná, a ve valenční struktuře slovesa pak v očekávané valenční pozici chybí. Zde uvedeme pouze poměrně jednoduchý případ "zrecipročnění" Aktoru a Patientu v (11); mohou však nastat i případy složitější.

(11) Jan a Marie se líbají (=Jan líbá Marii (a zároveň) Marie líbá Jana)

- byly vypracovány pracovní zásady pro tvorbu valenčních rámců sloves a od nich odvozených substantiv a adjektiv pro anotátory při anotaci

- je připravován mezinárodní projekt (účast lingvistů z ÚJČ AV ČR, UK, MU Brno, UP Olomouc a univerzit v Tübingen, Sheffieldu a Neapoli) ´Gramatika češtiny v českém národním korpusu´, jehož cílem je získat materiálovou bázi pro velkou mluvnici současné psané češtiny, v níž budou gramatické jevy dokumentovány jejich skutečným užíváním v reprezentativních psaných textech publicistických, beletristických i odborných. U konkurenčních a periferních či méně centrálních struktur budou získávány údaje o frekvenci, textové distribuci a z toho plynoucím komunikativním statusu těchto struktur v současné psané komunikaci. Tento výzkum bude propojen jak s probíhající prací na desambiguaci morfologického značkování (vedenou V. Petkevičem), tak s budováním Pražského závislostního korpusu a jeho značkováním tektogramatickým (sémantickosyntaktickým), vedeným J. Panevovou.

- průběžně je doplňována databáze jazykové poradny, zejména o jevy, které jsou v dlouhodobém vývojovém pohybu, na které se tazatelé nejčastěji ptají, a ty, které představují nějaký teoretický problém.

(B) UPLATNĚNÍ VELKÝCH SOUBORŮ DAT PRO JAZYKOVOU ANALÝZU

- **CD "Prague Dependency Treebank 1.0"** (Pražský závislostní korpus - PZK) - kat. číslo LDC2001T10 (ref. Hajič, J. et al., 2001); http://ufal.mff.cuni.cz/pdt - bylo vydáno v září 2001 organizací Linguistic Data Consortium, Univ. of Pennsylvannia, Philadelphia, PA, USA, pro členy i nečleny LDC. Příprava k vydání CD v sobě zahrnovala následující dílčí činnosti:
  - příprava a realizace formálně-kontrolního a testovacího balíku programů pro výslednou podobu CD, provedení kontroly v několika fázích, zanesení oprav a změn vynucených kontrolou
  - zpracování dokumentace v angličtině
  - příprava a realizace licenčních smluv s dodavateli dat pro jejich bezplatné šíření
  - příprava a realizace grafiky vlastního produktu (CD, obal, leták)
  - příprava vlastního obrazu CD (image) se zajištěním plné meziplatformové kompatibility (Linux, Unix, MS Windows, Macintosh), provedení příslušných změn
  - příprava lokálních kopií, rozeslání partnerům k testování
- **vývoj a zpřesňování specifikací systému anotací na tektogramatické rovině**
  - byly vytvořeny **čtyři podúrovně anotování** na tektogramatické rovině a dvě fáze automatického před- a **post-zpracování**, a to v tomto pořadí:
    - automatické **předzpracování analytické roviny**, výsledkem je výchozí podoba dat pro ruční anotaci tzv. velkého souboru
    - ruční **anotování velkého souboru** (anotace struktury a funktorů)
    - ruční **anotování aktuálního členění** (hloubkový slovosled a hodnoty atributů pro aktuální členění) ve velkém souboru
    - ruční **anotování koreferencí** ve velkém souboru
    - automatické post-zpracování po skončení anotování tzv. velkého souboru, výsledkem jsou **data** připravená **pro anotování v** tzv. **vzorovém souboru**
    - ruční **anotace** tzv. **vzorového souboru**.
  - byla vypracována "**víceprůchodová" strategie anotace velkého souboru**: soubor se v jednom průchodu anotuje podle stávajících instrukcí bez ohledu na jejich případné průběžné změny až do dokončení jednoho a začátek dalšího průchodu
  - pokračovalo se ve vývoji specifikací valenčních rámců
  - vývoj souboru anotačních instrukcí (manuálu) pro anotátory
  - vypracovány pracovní zásady pro tvorbu valenčních rámců sloves, substantiv a adjektiv rámců anotátory při anotaci

- **manuální anotace na tektogramatické rovině**
  - během manuální anotace tektogramatické roviny probíhá komplexní testování software, anotačních pokynů a souběžný výzkum problematických jevů na všech 4 anotačních podúrovních
  - byl připraven **referenční soubor 1500 vět** na podúrovni velkého souboru
  - pro podúroveň velkého souboru a aktuálního členění, a dalšího pro anotaci polysémie pro příští rok bylo **vyškoleno** dalších **pěti anotátorů**
  - příprava dat pro první anotační průchod (tj. zpracování kompletní sady dat určených pro anotování na tektogramatické rovině v objemu 55 tis. vět)
  - zahájení prvního anotačního průchodu, do konce roku anotováno 12000 vět na podúrovni velkého souboru
- **vývoj valenčního slovníku**

Vedle pracovní verze valenčních rámců pro anotátory byla vytvořena **koncepce** detailně zpracovaného **elektronického slovníku slovesné valence**, která využívá teoretického rámce pro valenci vypracovaného uvnitř funkčního generativního popisu, elektronického sběru dat ze SSJČ a doposud dostupné části slovníkové databáze češtiny EuroWordNet (ve spolupráci s FI MU v Brně). Do tohoto slovníku se ukládají lexémy podrobně tříděné nejen podle valenčních rámců, ale i podle lexikálních významů. Specifikace lexikálních významů se opírá o primárnost a přenesenost významu, o sémantické hodnocení jednotlivých argumentů slovesa, o jejich morfologické vyjádření, o synset popisující lexém v bázi EuroWordNet. Vedle valenčních členů obligatorních a fakultativních v terminologii FGP se zaznamenávají také členy kvazivalenční a obvyklé (které mnohdy dospecifikovávají lexikální význam).

V r. 2001 byl ve slovníku zpracován vzorek cca 350 nejčastějších českých sloves (jako kritérium pro výběr sloužila četnost slovesa v Pražském závislostním korpusu). Průměrný počet rámců na sloveso v tomto vzorku je 2,9, průměrná velikost rámce (počet prvků rámce) je 3,0. U těch sloves, která už jsou v české verzi EuroWordNetu (z uvedeného vzorku to je zhruba polovina), byl k valenčnímu rámci zaznamenán i odkaz na příslušný synset nebo skupinu synsetů.

- vypracování tzv. anotačního valenčního slovníku na základě referenční sady anotovaných souborů (rozsah cca 1500 hesel)
- specifikace jednotného formátu uložení dat pro výsledný jednotný valenční slovník
- vybudování kombinované verze valenčního slovníku pro přímé použití při anotacích

- na základě materiálové sondy z PZK byla zkoumána nosnost lexikálního přístupu ke gramatice, a to zejména při zkoumání slovních tříd s velmi nízkými frekvencemi v korpusu. Výsledky byly publikovány ve stati (Uhlířová, 2001).

(C) METODOLOGIE A KOMBINACE RŮZNÝCH PŘÍSTUPŮ K JAZYKOVÉMU MODELOVÁNÍ

- vývoj **modulu pro zpracování neznámých slov** na základě statistických dat získaných z morfologické úrovně PZK 1.0.

- pokračování vývoje metod **morfologického značkování** na základě **loglineárního modelu** a výběru rysů; natrénování nového modelu na výsledných datech z PZK 1.0.
- **označkování ČNK** (duben 2001 - původní model, prosinec 2001 - nový model) pomocí nově vytvořených modelů; porovnání úspěšnosti předchozích a nových modelů.
- zajištění návaznosti trénovacích algoritmů a programů na systém redukce víceznačnosti založený na pravidlech vyvíjený na jiných pracovištích (pro systém značkování založený na kombinaci symbolického a statistického přístupu); provedení experimentů s první verzí systému, vyhodnocení s pozitivním výsledkem.

(D) MATEMATICKÉ A KOMPUTAČNÍ ZÁKLADY

- **softwarová podpora anotování PZK na tektogramatické rovině**:
  - **pokračování** ve **vývoji nástroje TrEd** (Tree Editor). Od dubna 2001 TrEd nahradil i v predběžné fázi testování původní nástroj GRAPH, který fungoval jen na platformě Windows (TrEd je platformově nezávislý, testuje se pro Linux a MS Windows). TrEd pracuje s původním formátem .FS i novým SGML (csts.dtd). Programování je možné jak původním makrojazykem GRAPHu (pomocí konverze), tak přímo v programovacím jazyce Perl. Program byl rovněž otevřen pro integraci lokálně spouštěného software (viz dále, přiřazování funktorů) i pro spolupráci s jiným software typu klient - server.
  - pokračování ve vývoji **automatické** procedury **předzpracování dat z analytické roviny** pro anotování.
  - **integrace** programu pro automatické robustní **přiřazení funktorů** v dokončené tektogramatické struktuře **do** editoru **TrEd**. Lze použít před i po ručním zpracování (vytvoření) struktury stromu na tektogramatické rovině, a to vždy v režimu "online" (při vlastní anotaci). Program doplněn o grafickou indikaci jistoty přiřazení pro usnadnění vizuální kontroly.
  - vývoj modulu pro TrEd umožňujícího formalizovanou tvorbu a použití valenčních rámců v průběhu anotace. Účelem modulu je formální kontrola v průběhu i po skončení ruční fáze anotace a usnadnění poanotačních změn v anotaci valencí. Vývoj modulu bude dále pokračovat, mj. i směrem k přímé podpoře manuální anotace.

(E) ROZPOZNÁVÁNÍ ŘEČI

V roce 2001 se CKL podařilo získat na řešitelská pracoviště MFF UK a ZČU dekodér pro rozpoznávání řeči od firmy AT&T. Tento softwarový nástroj využívá principy teorie konečných automatů. Všechny složky systému pro rozpoznávání řeči (akustický model, jazykový model, výslovnostní slovník) proto musí být převedeny do podoby konečného automatu, což je na jedné straně mnohdy netriviální úloha, na druhé straně však tento formalismus umožňuje jednotný přístup ke všem komponentám systému a také zaručuje snadnou integraci dalších modulů. Studiu dekodéru byla proto v roce 2001 věnována značná pozornost.

V průběhu roku 2001 byly s využitím výše zmíněného dekodéru provedeny experimenty s rozpoznáváním souvislé mluvené řeči v úloze s velkým slovníkem (několik desítek tisíc slov). V těchto experimentech byly testovány 2 různé druhy jazykových modelů – standardní *n*-gramové modely založené na slovech a *n*-gramové modely založené na kmenech a koncovkách (morfémech). V obou případech byl nejprve použit bigramový

jazykový model a AT&T dekodérem vygenerovány seznamy *N*-nejpravděpodobnějších posloupností slov (tzv. *N*-best listy). Ty pak byly reskórovány pomocí trigramových modelů; pro morfémy byla použita speciální forma modelu, která zohledňuje odlišnosti v predikci kmene a koncovky. Přestože morfémové modely vykazovaly při stejné velikosti slovníku mnohem nižší procento slov mimo slovník (tzv. *OOV rate*), jejich úspěšnost při rozpoznávání byla mírně horší než v případě standardních modelů založených na slovech. Tato skutečnost je způsobena především nižší kvalitou bigramového morfémového modelu, který byl použit při generování *N*-best listů. Výsledky těchto experimentů byly publikovány v (Psutka, 2001).

Na snížení *OOV rate* je zaměřen také další přístup, který byl v roce 2001 zkoumán. Jedná se o metodu dvouprůchodového rozpoznávání, kdy po prvním průchodu je na základě analýzy výsledků rozpoznávání adaptován slovník, zkonstruován nový jazykový model a ten je pak použit v druhém průchodu. Analýza výsledků rozpoznávání opět využívá rozkladu slov na kmeny a koncovky. Touto metodou, jež byly publikována v (Ircing, 2001a) se podařilo dosáhnout slibných výsledků, které bude třeba ověřit v dalším období na větší množství dat.

Teoretický rozbor problematiky jazykového modelování češtiny a vysoce flexivních jazyků obecně je také obsažen v práci ke státní doktorské zkoušce (Ircing, 2001b), která byla obhájena v září 2001.

Pracovníci CKL se též podíleli na sběru a anotaci rozhlasových zpráv vysílaných zahraničním vysíláním Rádia Praha v pěti jazycích (čeština, němčina, angličtina, francouzština a španělština). Připravený korpus paralelních řečových nahrávek a textových přepisů byl jedním ze zdrojů pro úvodní experimenty, jež proběhly na Johns Hopkins University v Baltimore při návrhu automatického vícejazyčného systému s cílem rozpoznat a porovnat informace a znalosti obsažené ve vícejazyčných řečových signálech. První výsledky experimentů byly publikovány v (Jelinek, 2011).

Na půdě Centra probíhaly v první polovině roku 2001 též některé dílčí práce končícího tříletého projektu ME293 „Speech Recognition of a Slavic Language: Czech“. Tento společný projekt JHU Baltimore, MFF UK, ZČU v Plzni a TU v Liberci byl podporován MŠMT především z hlediska výměny pracovníků (dlouhodobé pracovní pobyty) a přinesl mnoho cenných výsledků, na něž bude v následujícím období dále navázáno právě v Centru. Závěrečné výsledky projektu ME293 byly publikovány na prestižní konferenci EUROSPEECH’2001 v Aalborgu.

V oblasti zkoumání větné prozodie pro účely analýzy větného členění, a to zejména z hlediska možnosti charakterizovat rozdíly mezi různými typy přízvuku, bylo hlavní úsilí v průběhu roku 2001 soustředěno na tvorbu prosodického korpusu spontánní řeči (Peterek, 2001). Byly průběžně nahrávány spontánní dialogy z televizních pořadů, které byly následně anotovány a opatřovány prozodickými značkami. V prozodickém značkování se vycházelo ze systému *Tilt*, kde jsou značkovány prozodické události typu větný důraz, klesavý či stoupavý hraniční tón a jejich kombinace. Spojité parametry těchto událostí (číselná vyjádření průběhu jejich základního hlasivkového tónu) jsou automaticky počítány softwarovým nástrojem EST (Edinburgh Speech Tools). Takto označkovaný korpus má posloužit jak k natrénování automatického detektoru prozodických událostí, tak i ke generování intonační křivky, což jsou důležité prostředky studia větné prozodie. Kromě prozodických událostí jsou označovány i neřečové události ovlivňující průběh dialogu, například slyšitelné nádechy, smích, zvukově vyjádřená potvrzování a nesouhlasy. V současnosti je prozodicky označkováno kolem

čtyř hodin dialogů. Započaly též práce s nahráváním dialogových úloh typu MapTask a čtených prozodicky charakteristických vět. Ke zpracování dostupných dat byly vytvořeny programy schopné automaticky extrahovat z prozodického přepisu dialogu jednotlivé promluvy, generovat jejich fonetické přepisy, nalézt umístění jednotlivých fonémů v čase (pomocí HTK softwaru) a pak dále analyzovat průměrné hodnoty časového trvání fonémů, jejich dynamiky i základního hlasivkového tónu. Takto zpracované promluvy jsou vhodné i k vizuální analýze prováděné lingvisty, kteří tak mohou podrobně studovat průběhy intonačních a dynamických křivek pomocí dalších nástrojů (např. programem Wavesurfer).

(F) VÍCEJAZYČNÉ ZDROJE, STROJOVÝ PŘEKLAD a APLIKACE

- paralelní česko-anglický publicistický korpus vytvořený z textů anglické a české verze časopisu **Reader's Digest** Výběr byl v rozsahu 50000 paralelních vět **zařazen na CD-ROM PDT 1.0**. Korpus byl na české i anglické straně morfologicky označkován, česká část byla navíc automatickým parserem doplněna o syntaktické závislosti.
- výzkumný tým zabývající se problematikou **strojového překladu** provedl experiment **obohacené překladové paměti o kratší segmenty**. Tento projekt byl zaměřen na automatické získávání vzájemně si odpovídajících segmentů z relativně malých paralelních korpusů (3000 až 10000 vět). Navržený postup měl být zcela automatický a jazykově nezávislý, proto byly použity statistické metody. Jazykový model, konkrétně hodnoty vzájemné informace, byly indikátorem nastavení hranic segmentů. Základem pro navržení skórovací funkce pro hodnocení hypotéz párování segmentů byl IBM překladový model 2 (Brown et al., 1993). Nejlepší párování segmentů bylo vybráno pomocí A* algoritmu, výsledkem byl překladový slovník segmentů. Další experimenty budou spočívat v začleňování jazykově závislých pravidel pro získávání jednojazyčné segmentace a v implementaci grafického uživatelského rozhraní.
- Byla vypracována specifikace projektu **vícejazyčného strojového překladu** na základě analýzy/syntézy do/z tektogramatické roviny a strukturního transferu. Anotace dat bude zahájena v příštím roce, nutný překlad dat pro získání vzorků paralelního korpusu se zahajuje ke konci r. 2001 (viz dále). Pro úspěšné pokračování v tomto směru výzkumu byly zahájeny tyto práce:
  - jednání o využití dat vyvíjených na University of Pennsylvannia („Proposition Bank") pro účely strojového překladu
  - zahájení **překladu** (tvorba terminologického glosáře) z angličtiny **do češtiny** části dat z **Penn Treebank v.** 3
  - experimenty s **porovnáváním tektogramatické reprezentace českých a anglických** vět na vzorku paralelního korpusu (Reader's Digest)
  - experimenty s **generováním češtiny z tektogramatického zápisu** jako předstupeň k obdobnému experimentu s angličtinou (připravováno pro příští rok)
  - **úprava programovacích nástrojů** pro manuální anotaci směrem k **jazykové nezávislosti**, a to i z hlediska jazyků se zcela odlišnými znakovými soubory a systémem psaní
- byly vypracovány **zásady pro anotaci arabštiny na morfologické a analytické rovině**

Projekt automatického překladu z arabštiny do angličtiny se soustřeďuje na ekonomický jazyk. Stávající korpus arabštiny CLARA (Corpus Linguae Arabicae - 67 mil. slov), ani LDC Arabic Corpus (asi 100 mil. slov) nejsou pro statistické trénování použitelné, jelikož neobsahují anglický protějšek textu ani lingvistickou anotaci. Požadovaný rozsah paralelního korpusu je "co nejvíce" slov, min. 1 mil. slov. Zdrojem jsou výroční ekonomické zprávy Ligy arabských států, OSN, Světové banky, MMF apod., přes Internet lze však získat jen malou část dat.

Morfologickou analýzu arabštiny poskytne dvouúrovňový FST systém od firmy Xerox, příp. morfologická analýza vytvářená v LDC. Disambiguace se provede zpočátku ručně na standardně velkém subkorpusu (300 tis. až 1 mil. slov), rovněž tak syntaktická a tektogramatická analýza. Na těchto datech budou trénovány statistické analyzátory na všech třech úrovních. Pro angličtinu ruční práce odpadá, k dispozici jsou strukturovaná data z Penn Treebank, která budou dále anotována buď manuálně na tektogramatické rovině, nebo bude převzat tzv. Proposition Bank z Univ. of Pennsylvannia, který bude do tektogramatické roviny zkonvertován.

Zatímco všechny tři úrovně značkování PZK mohly probíhat nezávisle na sobě, arabština vyžaduje primární provedeni morfologické disambiguace. Jediný arabský řetězec může totiž obsahovat jak řídící slovo, tak i připojenou předložku, zájmeno, částici apod. Rozklad na jednotlivá slova, resp. větné členy, je stejně obtížný a nejednoznačný jako morfologická analýza a probíhá současně s ní.

Z pragmatických důvodů vyvíjíme poziční systém morfologických značek, který je, stejně jako výstup morfologické analýzy, postaven na arabských gramatických kategoriích. Při syntaktickém rozboru je třeba řešit zvláštní arabské konstrukce, jaké představují jmenná věta bez sponového slovesa, slovesa, jmenné spojky či fázová slovesa.

K anotaci používáme programy DA a TrEd (který byl kvůli pravo-levému směru psaní a výstavby stromu mírně upraven). Na anotacích se přímo podílí pět studentů, práce na arabské části je koordinována dalšími dvěma pracovníky.

- projekt **Česílko**: v rámci projektu automatického překladu mezi slovanskými jazyky Česílko byly v minulém roce provedeny tyto práce:
  - byl úspěšně proveden zevrubný test pro jazykový čeština-slovenština v rozsahu 50 000 slov na datech z encyklopedie Diderot
  - v úzké návaznosti na provedený test byly doplněny slovníky systému, zejména překladový česko-slovenský slovník a slovník používaný pro generování slovenštiny
  - na datech z oblasti manuálů k podnikovým informačním systémům byl proveden test jazykového páru čeština-polština, který prokázal relativně dobrou kvalitu překladu (úspěšnost cca 70%)
  - byla provedena analýza typů chyb překladu mezi češtinou a polštinou z hlediska jejich odstranitelnosti v další fázi vývoje systému
- aplikace lingvistických metod při vyhledávání informací v českých textech

Pro podporu výzkumu a provádění experimentů jsme vyvinuli **experimentální lingvistickou databázi** MATES. V současné době je v databázi uloženo **přes 40 tisíc českých textových dokumentů** a je možné statisticky vyhodnocovat distribuci nejrůznějších lingvistických fenoménů v nich. Dílčích úloh, které systém MATES využívají nebo mohou využít, je mnoho. Rozpracované jsou zejména charakterizace sémantického obsahu

dokumentů pomocí hierarchie sémantických konceptů, sémantická klasifikace lexikálních jednotek, disambiguace slovních významů, nástroje pro automatickou detekci sémanticky signifikantních kolokací. Na některých úlohách také významnou měrou participují studenti vyšších ročníků, diplomanti a doktorandi. Zároveň pracujeme na vývoji nové verze systému MATES, která v dohledné době přinese jednak zefektivnění některých výpočetně náročných operací, jednak rozšíření funkcionality systému.

2. Personální a organizační zabezpečení činnosti Centra

UK MFF

Složení pracovního týmu z hlediska kvalifikace ve vztahu k pracovní náplni v Centru je vyvážené. Na řešení projektu se podílí celkem 34 pracovníci (v roce 2000: 38) – 23 (v roce 2000: 18) pracovníci s úvazkem rovným nebo vyšším než 0,7 a 11 (v roce 2000: 20) pracovníků s úvazkem nižším než 0,7. Rovněž z hlediska věku pracovníků je 24 pracovníků mladších než 35 let (jejich kapacitní podíl na projektu je 70%). Prof. Frederick Jelinek, Dr. h.c. byl po celý rok 2001 hostujícím profesorem MFF UK.

ZČU

V personálním obsazení plzeňské sekce Centra nedošlo během roku 2001 k žádné změně, tj. v Centru pracovali čtyři pracovníci s celkovým úvazkem 1,4: Josef Psutka (0,3), Pavel Ircing (0,7) a dva spolupracující studenti (0,2). Plzeňská sekce sídlí v prostorách katedry kybernetiky FAV, ZČU s tím, že může dle potřeby využívat veškeré technické a personální zabezpečení katedry a univerzity.

ÚJČ AV

Vzhledem k delší nemoci PhDr. L. Uhlířové, Csc. byl po vzájemné dohodě jmenován vedoucím skupiny CKL v ÚJČ PhDr. František Štícha, CSc.

V oddělení gramatiky a jazykové kultury ÚJČ na částečný úvazek pracovali na poradenských úkolech a zároveň na vyhledávání styčných bodů mezi těmito úkoly a budováním Pražského závislostního korpusu studentky Kamila Stejskalová, Eva Flanderková a Dagmar Vlčková. Od 1. 11. 2001 byla na tyto úkoly přijata na celý úvazek Markéta Pravdová. Ta se bude soustavněji věnovat možnému usouvztažňováním poradenských problémů z oblasti syntaxe s některými otevřenými otázkami tektogramatického značkování Pražského závislostního korpusu.

3. Spolupráce centra

Pracovníci Centra úspěšně rozvíjeli již existující a navázali kontakty jak s tuzemskými, tak zahraničními pracovišti a projekty. Svědčí o tom jak aktivní účast na mezinárodních konferencích a konzultace s předními zahraničními badateli (viz body 5(b) a 6 této zprávy), tak i konkrétní zapojení do mezinárodních struktur a akcí, z nichž uvádíme několik nejpodstatnějších:

a) Při příležitosti 34. výročního zasedání Evropské lingvistické společnosti (Societas linguistica Europaea, SLE) byla **Eva Hajičová** pověřena uspořádáním a vedením workshopu na téma ´Empirical Methods in the New Millenium - Linguistically Interpreted Corpora´. Toto pozvání bezprostředně vyplývalo z toho, že Pražský závislostní korpus, jehož budování patří k základnímu vědeckému projektu Centra, je považován za nejpokročilejší evropský projekt lingvistického zpracování (značkování) korpusu textů.

Na workshopu přednesla referát o tomto projektu pracovnice CKL **V. Řezníčková** a zúčastnili se ho i z ÚJČ **L. Uhlířová** a **F. Štícha** (s vlastním referátem (Štícha, 2001), ve kterém šlo o poukaz na bezprecedentní možnosti, které označkovaný korpus poskytuje empirickému výzkumu syntaxe).

b) Na základě pozvání Mezinárodního komitétu lingvistů (CIPL) bylo Centrum komputační lingvistiky pověřeno pořádáním XVII. Mezinárodního kongresu lingvistů v Praze v r. 2003 (http://www.cil17.org). **Předsedkyní** organizačního výboru byla jmenována **Eva Hajičová**, **sekretářkou** výboru je pracovnice CKL **A. Kotěšovcová**. Kongres se bude konat ve dnech 24.-29. července 2003 v Kongresovém paláci v Praze a očekává se účast více než 1000 účastníků. Jde o velmi prestižní záležitost, kongresy se konají v pětiletém odstupu a jsou vrcholným setkáním lingvistů celého světa (v roce 1992 byl kongres v Quebeku, v r. 1997 v Paříži).

c) Pracoviště ÚJČ zahájilo spolupráci s prof. R. Köhlerem (Universität Trier). Cílem spolupráce bude zjišťovat statistické charakteristiky češtiny na základě korpusového materiálu opatřeného gramatickými anotacemi a porovnávat vlastnosti gramatických popisů jazyků (koncepce bezprostředních složek a závislostní koncepce). Prof. Köhler navštívil Prahu a přednesl v CKL přednášku.

d) V rámci přípravy projektu ´Gramatika češtiny v Českém národním korpusu´ existuje úzká spolupráce vedoucího skupiny Centra v ÚJČ s Českým národním korpusem. Další spolupráce se rozvíjí s MU Brno (prof. Petr Karlík), UP Olomouc (doc. M. Hirschová) a univerzitami v Tübingen (prof. T. Berger), Sheffieldu (N. Bermel) a Neapoli (prof. F. Esvan).

e) Pracovníci CKL jsou členy hlavních výborů či funkcionáři reprezentativních mezinárodních vědeckých společností oboru: International Committee of Computational Linguistics (**E. Hajičová** – místopředsedkyně, **P. Sgall** – člen výboru), International Speech Associtaion (**E. Hajičová** – členka výkonného výboru).

f) Pracoviště je zapojeno do mezinárodního projektu TEI (Text Encoding Initiate); členem výkonného výboru je **J. Hajič.**

4. *Podpora a výchova mladých výzkumných pracovníků*

Obě sekce CKL, jejichž mateřským pracovištěm je vysoká škola, jsou výraznou měrou zapojeny do doktorských programů a vychovávají mladé výzkumné pracovníky, jimž dávají příležitost k vlastnímu bádání i jeho prezentaci

doktorské studijní programy

CKL se významnou měrou podílí na doktorském programu I3 - Informatika - Počítačová lingvistika na Matematicko-fyzikální fakultě UK a na doktorském programu v oboru Kybernetika na Fakultě aplikovaných věd, Západočeská univerzita.

V letošním roce úspěšně dokončili svá doktorská studia 4 doktorandi školitelky Jarmily Panevové, z toho dva zaměstnanci Centra – Markéta Straňáková-Lopatková a Vladislav Kuboň.

Školiteli nebo konzultanty 6 nově přijatých studentů doktorského studia na MFF UK se stali pracovníci Centra Eva Hajičová, Jarmila Panevová, Jan Hajič a Barbora Vidová Hladká.

Na úkolech plzeňské sekce Centra se podílí několik studentů, kteří jsou za svou práci odměňování stipendiem. Většinou vykonávají pomocné práce týkající se anotací a transkripcí řečových nahrávek, přípravy slovníků ap. Josef Psutka je školitelem 6 doktorandů (tito doktorandi studují obor Kybernetika akreditovaný na Katedře kybernetiky ZČU).

Kromě vlastní vědecké práce (příp. práce na projektech) se doktorandi aktivně podílejí na akcích pořádaných pod hlavičkou Centra. **Konkrétní akce** roku **2001** spolu s přehledem doktorandů aktivních v té které akci jsou uvedeny v **Příloze A** na konci Zprávy.

5. Způsob zpřístupnění výsledků a výstupů Centra veřejnosti

Výzkum v rámci projektu byl podstatnou měrou prezentován laické i odborné veřejnosti, a to jak cílenými akcemi, tak i aktivní účastí pracovníků Centra na mezinárodních konferencích. Kompletní **seznam** všech publikovaných či přednesených **vědeckých statí** je uveden v **Příloze B** na konci Zprávy.

6. Cílené akce

a) Byla nastavena internetovská stránka http://ckl.mff.cuni.cz , která poskytuje dokonalý přehled všech aktivit, které se k Centru vztahují.

b) Ústav Českého národního korpusu vydal na konci roku CD se vzorkem dat z korpusu Synset2000. Střední školy by měly představovat skupinu uživatelů, pro které je chystané CD určeno. Protože PZK představuje dokonalou nadstavbu nad částí dat z národního korpusu, dodali jsme na CD upoutávku v podobě textu s názvem "Pražský závislostní korpus aneb Hledáme nové možnosti pro mluvnický rozbor textu". Formou blízkou studentům středních škol jsme se snažili zdůraznit možnosti, které anotovaný korpus přináší do světa výuky českého jazyka.

c) V letošním roce byl připraven k vydání první díl (**autoři**: Eva Hajičová, Jarmila Panevová, Petr Sgall) (v tisku nakl. Karolinum) dvoudílných skript **Úvod do teoretické a počítačové lingvistiky.**

d) Výsledky práce pracoviště ÚJČ AV jsou zprostředkovaně prezentovány jazykovým poradenstvím.

e) V rámci Dne otevřených dveří UK MFF, 29. listopadu 2001, proběhl v Národním domě na Vinohradech i Den otevřených dveří CKL. Zájemci, především z řad středoškolských studentů, měli možnost zhlédnout prezentaci programových prostředků používaných při vytváření a anotování Pražského závislostního korpusu a získat informace i o dalších projektech probíhajících v rámci CKL. Část návštěvníků projevila o činnost CKL hlubší zájem a vyžádala si detaily, které jim poskytli **Jiří Havelka** a **Jan Štěpánek**, kteří CKL prezentovali v dopolední části. Na odpolední prezentaci informatické sekce přednesl **Jan Cuřín** referát na téma strojový překlad mezi přirozenými jazyky.

Název projektu : *Centrum komputační lingvistiky*
Řešitel: Prof. PhDr. Eva Hajičová, DrSc.
Příjemce: 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. 2002[2]

Uvádíme zde upřesnění cílů pro r. 2002, 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 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). Nově je zařazena významná organizační aktivita Centra, totiž příprava 17. Světového kongresu lingvistů

V následujícím přehledu jsou jednotlivé body konkretizovaného programu pro rok 2002 označeny (T1) až (T6), popř. dalším členěním, a to v souladu s časovým harmonogramem uvedeným pod přehledem.

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

Pražský závislostní korpus (PZK) je stěžejním projektem CKL v roce 2002. PZK bude rozšiřován na tektogramatické rovině o anotaci struktury, aktuálního členění a koreference. Do anotace bude z důvodů konzistence integrován valenční slovník, který se jinak připravuje separátně. Předpokládáme, že bude anotováno alespoň 10,000 vět na tektogramatické rovině. Dále bude podle potřeby probíhat manuální anotace lexikálně-sémantická (rozlišení polysémie).

Bude dále rozvíjeno softwarové vybavení pro anotace, následnou kontrolu a zpracování dat a dokumentace obecně.

Z hlediska souvisejících prací, příp. dokončení specifikace v anotačních pokynech bude třeba pracovat zejména na:

T1-1 specifikaci valenčních rámců neslovesných slovních druhů,

T1-2 verifikaci stanovených kritérií pro přiřazování valenčních rámců těmto slovním druhům na základě PZK,

T1-3 verifikaci a případné upřesnění pravidel pro přiřazování hodnot atributu pro zachycení aktuálního členění na základě korpusových dat,

T1-4 upřesnění pravidel pro rekonstrukci uzlů elidovaných v povrchové podobě věty, co nejvyšší stupeň jejich algoritmizace a implementace,

T1-5 prohloubení koncepce valenčního slovníku a její uplatnění na vytvoření několika set slovníkových hesel ověřených na datech z PZK

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

T1-6 Pro rozšíření množství dat na morfologické úrovni bude provedeno dokončení konverze tzv. ÚJČ korpusu (600 tis. slov morfologicky označkovaných v 70. letech v ÚJČ) do "PZK - formátu".

T1-7 Manuální anotace PZK

T2: Strojový překlad (bod F původního návrhu)

Projekt strojového překladu bude fakticky zahájen v r. 2002. Po důkladné analýze problému v 2. pol. 2001 bylo rozhodnuto založit systém strojového překladu na těchto zásádách:

1. Systém bude založen na „klasickém“ modelu ANALÝZA – TRANSFER – SYNTÉZA, přičemž jednotkou překladu bude rovněž tradičně jedna věta.

2. Pro reprezentaci struktury a významu věty na úrovni transferu bude použita tzv. tektogramatická rovina, použitá rovněž v Pražském závislostním korpusu.

3. Jednotlivé komponenty systému budou založeny na použití převážně statistických metod, vhodně doplněných existujícími lexikálními databázemi.

4. Projekt bude pokud možno vícejazyčný s jazyky rozdílného typu; první dvojice bude překlad z češtiny do angličtiny.

Projekt bude rovněž těžit z předchozích projektů v oblasti strojového překladu řešených nebo spoluřešených v ÚFALu a jeho předchůdcích (zejména ze systému anglicko-českého překladu APAČ a česko-ruského překladu RUSLAN), a využívat nástrojů a lingvistických dat vytvořených v rámci bývalé Laboratoře pro zpracování jazykových dat ÚFAL a CKL, zejména pak bude velmi těsně navazovat na projekt Pražského závislostního korpusu, jehož náročně vytvořená data tak bude ve velké míře využívat, neboť projekt strojového překladu v sobě zahrnuje téměř všechny problémy zpracování přirozeného jazyka, od morfologie přes syntax a sémantiku až po generování (syntézu).

Projekt bude konzultován s pracovišti zabývajícími se strojovým překladem nabo reprezentací věty v angličtině a analýzou angličtiny v zahraničí (University of Pennsylvannia, ISI, Johns Hopkins University). Přepokládáme, že nejtěsněji budeme spolupracovat s Center for Language and Speech Processing na Johns Hopkins University v Baltimore, společně budeme usilovat o přídavné prostředky na začlenění dalších jazyků (především arabštiny). Pokud se podaří zajistit dostatek prostředků, budeme pokračovat ve spolupráci s FF UK, Katedrou Předního a Blízkého Východu, se kterými bychom další práci na projektu koordinovali na základě našich zkušeností s Pražským závislostním korpusem i se strojovým překladem obecně.

T3: Zpracování mluvené řeči (bod E původního návrhu)

T3-1 V následujícím roce se zaměříme především na vývoj nástrojů, které by umožnily převod nestandardních forem jazykových modelů do formátu konečných automatů. Tyto nástroje jsou nezbytné především pro důkladné otestování možností morfémových jazykových modelů zakomponovaných přímo do AT&T dekodéru. Samozřejmě budou dále probíhat experimenty s rozpoznáváním souvislé řeči v úloze

s velkým slovníkem. Budou hledány postupy, jak dále snížit *OOV rate* a zajistit další zvýšení přesnosti rozpoznávání.

Pro několik následujících let počítáme s tím, že pracovníci Centra, zabývající se problematikou rozpoznávání souvislé řeči a problematikou statistického modelování jazyka, budou participovat na řešení mimořádně rozsáhlého projektu „MALACH“ (Multilingual Access to Large Spoken Archives) podporovaného National Science Foundation (USA), Project #0122466 (www.clsp.jhu.edu/research/malach). Tento vysoce prestižní projekt byl přijat na období 5 let a jeho cílem je vývoj systémů pro automatický přepis svědeckých výpovědí lidí, kteří přežili holocaust. Na projektu participují dále Visual History Foundation v Hollywoodu, Johns Hopkins University v Baltimore, University of Maryland, IBM, MFF UK v Praze a ZČU v Plzni. Během následujících pěti let má být zpracováno minimálně 5 jazyků střední a východní Evropy. Vedle přípravy řečových a jazykových korpusů pro trénování akustických a jazykových modelů by měly MFF UK a ZČU spolupracovat i na přípravě odpovídajících systémů rozpoznávání řeči. Počítá se s tím, že na akustické analýze a modelování svědeckých výpovědí by se podíleli zejména pracovníci ZČU v rámci Výzkumného záměru MSM235200004 (dlouholeté zkušenosti), kdežto na problematice jazykového modelování a na přípravě systému rozpoznávání řeči by participovali pracovníci CKL. Účast pracovníků Centra ve výše zmíněném projektu poskytuje nebývalou možnost vyzkoušet dosud vyvinuté techniky jazykového modelování na rozsáhlém souboru nahrávek spontánní řeči.

T3-2 Při studiu prozodie a jejího vztahu k aktuálnímu členění se soustředíme na natrénování prozodických modelů pomocí již přepsaných dat a budeme pracovat na vylepšení software sloužícího k vizualizaci prozodických parametrů.

T3-3 Rozšíříme prozodickou databázi o dialogy MapTask a čtené prozodicky charakteristické věty.

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 2002 bude výzkum pokračovat v následujících bodech:

T4-1 Valence slovesa i neslovesných slovních druhů v souvislosti s realizací či nerealizací členů valenčního rámce v povrchové podobě věty.

T4-2 Zachycení specifických jevů aktuálního členění věty, jako je negace a fokalizátory.

T4-3 Empirický výzkum větného kontrastu a začlenění jeho popisu do celkového rámce formálního popisu aktuálního členění.

T4-4 Studium jevů přesahujících rámec věty a patřících do oblasti diskursu, ovšem souvisejících s aktuálním členěním věty (zejména anaforické odkazování a koreference), a návrh na doplnění anotačního scénáře o zachycení těchto jevů v PZK.

T4-5 Specifikace reprezentace těch sémantických (kognitivních) aspektů, které přesahují jazykový význam, pro případné doplnění anotačního scénáře PZK o další úroveň.

T4-6 V oblasti matematických metod se budeme věnovat experimentům s modelováním pomocí maximální entropie, a to zejména vzhledem k problému morfologického značkování. Bude využito širší množiny možných rysů, a zároveň bude systém doplněn o iterativní výpočet optimálních hodnot parametrů pro vybrané rysy.

T5: Vyvíjení některých menších aplikačních systémů (bod F původního návrhu)

Vedle soustředěné práce na projektu strojového překladu uvedeného v bodě T2 výše budou pokračovat práce na dalších menších aplikačních systémech:

T5-1 Bude dále rozvíjen projekt Česílko, strojový překlad z češtiny do slovenštiny, jako jednoduchá aplikace některých metod zpracování přirozeného jazyka

T5-2 Budou provedeny nové experimenty s překladem do polštiny za použití podobně jednoduchých metod jako v projektu Česílko a bude vyhodnocena úspěšnost. Z podrobného vyhodnocení problémů bude rozhodnuto o případném začlenění dalších netriviálních modulů do systému česko-polského překladu.

T5-3 Budou pokračovat práce na vývoji systému pro vyhledávání informací se silnou podporou lingvistického výzkumu.

T6: Organizační přípravy 17. Světového kongresu lingvistů

Pracovníci Centra budou **pokračovat v organizačních přípravách 17. Světového kongresu lingvistů**, který se bude konat v Praze v červenci 2003; předsedkyní organizačního výboru je Eva Hajičová, jeho sekretářkou Anna Kotěšovcová, na práci organizačního výboru se podílí i řada dalších pracovníků Centra (Jarmila Panevová, Alena Böhmová, Kiril Ribarov i další). Vzhledem k tomu, že jde o velice prestižní světové setkání lingvistů (pořádané každých pět let na různých místech různých kontinentů, v Praze se dosud nekonalo), je jeho příprava i vlastní konání nesmírně důležitou aktivitou Centra s mezinárodním dosahem.

Časový harmonogram:

Měsíc I II III IV V VI VII VIII IX X XI XII

T1-1 o----------/
T1-2 o----------/
T1-3 o----------/
T1-4 o----------/
T1-5 o----------/
T1-6 o----------/
T1-7 o----------/
T2 o----------/
T3-1 o----------/
T3-2 o----------/
T3-3 o----------/
T4-1 o----------/
T4-2 o----------/
T4-3 o----------/
T4-4 o----------/
T4-5 o----------/
T4-6 o----------/
T5-1 o----------/
T5-2 o----------/
T5-3 o----------/
T6 o----------/

Další aktivity plánované pro rok 2002

- Budeme **pokračovat** ve vydávání **technických zpráv** (ve spolupráci s ÚFALem MFF UK) o dílčích výsledcích výzkumu; v roce 2002 předpokládáme vydání 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** (s některými jsou již dojednána data přednášek: prof. Emmon Bach a prof. C. Townsend z USA) a pořádání alespoň dvou několikatýdenních intenzivních přednáškových kursů zahraničních profesorů v průběhu kalendářního roku 2002, které se budou týkat aktuální problematiky řešené v rámci programu CKL (prof. Barbara Partee, UMass, Amherst, USA, jarní semestr, z oblasti formální sémantiky, a prof. F.. Jelineka, JHU, Baltimore, USA, z oblasti rozpoznávání mluvené řeči).
- Centrum se podstatnou měrou podílí a bude podílet na organizaci 17. běhu mezinárodních **cyklů přednášek Centra Viléma Mathesia** v Praze 11.-22. března 2002. Finanční podpora, která umožňuje nabídnout 30 stipendií i částečné úhrady cestovného pro studenty a uhradit cestovné a pobyt 12 zahraničním profesorům, je poskytována z grantu Evropské Unie (EuroSummerSchool) a z grantu Higher Education Support Programme, veškeré organizační zajištění je dílem pracovníků Centra. Bezplatně se přednášek zúčastní kolem 20 českých doktorandů a mladých vědeckých pracovníků.

PŘÍLOHA A *Akce CKL*

VÝJEZDNÍ SEMINÁŘE CKL

V průběhu roku 2001 se konaly dva semináře řešitelů a spoluřešitelů projektu (tzv. výjezdní semináře) v Rokytnici nad Jizerou. Hlavním cílem těchto seminářů bylo společné soustavné projednávání dílčích bodů z pracovní náplně Centra za účasti zástupců jednotlivých „spřátelených" pracovišť, i mimopražských.

21.1. – 26. ledna 2001, Rokytnice nad Jizerou

Semináře se zúčastnila většina pracovníků MFF UK , ZČU Plzeň „reprezentoval" Pavel Ircing a ÚJČ Ludmila Uhlířová. Ze zahraničních spřátelených pracovišť byly přítomny Marie-Anne Moreaux (INALCO, Paříž, Francie) a Sabine Doenninghaus (Univerzita v Basileji (Die Univarsität Basel), Švýcarsko). Během jednotlivých pracovních sezení byla diskutována následující témata:

- upřesňování **instrukcí pro anotátory** PZK:
  - problematika zavěšení rematizátorů
  - konstrukce se slovesem být a s příčestím trpným
  - problematické valenční rámce sloves a deverbativ
  - zachycení zvláštního typu gramatické koreference – „Control"
  - zachycení společného rozvití koordinace a elidovaných členů v koordinaci
  - problematika zpracování cizojazyčných frází v textu
- **srovnání prvních souborů** označkovaných různými anotátory:
  - hodnocení počátečních výsledků
  - vymezení hlavních diferencí ve značkování
- **anotátoři a jejich „softwarové vybavení / zázemí"**
  - hodnocení počátečních zkušeností s grafovým editorem
  - návrhy na další možnosti zdokonalování grafového editoru (makra)

20.9. – 25. září 2001, Rokytnice nad Jizerou

Personální složení bylo velmi podobné složení zimního semináře. Dále byli přítomni studenti FF UK, oboru bohemistika Jakub Dotlačil a Kateřina Součková, kteří se podílejí na tektogramatickém značkování PZK. Intenzivní diskuse byly věnovány následujícím aktivitám:

- projednávání poznámek k textům nových **dvoudílných skript** vydávaných ÚFAL (ve spolupráci s CKL)
  - 1. díl: Teoretická lingvistika – k vydání v roce 2001
  - 2. díl: Počítačové zpracování přirozeného jazyka – k vydání 2002, jasně byl formulován obsah jednotlivých kapitol spolu s jmenováním autorů zodpovědných za příslušné kapitoly
- zprávy o **spolupráci se zahraničními pracovišti**:
  - Ivona Kučerová: zpráva z dvoutýdenního pobytu na pracovištích v USA (Maryland a Philadelphia), možnosti spolupráce PZK a Proposition Bank
  - Jan Hajič: spolupráce s pracovišti v USA
    - překlady čeština, angličtina, arabština, čínština

- rozpoznávání mluvené řeči Baltimore
- přepis projevů obětí holocaustu

- upřesňování **instrukcí pro anotátory** PZK:
  - vymezení specifik **fázových sloves** oproti **modálním slovesům**; přiřazování valenčních rámců fázových sloves
  - zachycení **dispoziční modality**
  - vymezení **zájmenných lemmat**
  - zachycení **recipročních vztahů** mezi jednotlivými členy valenčních rámců **reciprok**
  - zachycení vztahu **přímé řeči** a uvozovacího slova
  - zachycení vztahu závislosti **časových a místních syntagmat** a řídícího slovesa
  - problematika **odkazovacích vět**
  - konstrukce se **slovesem být** a tradičním **doplňkem**

---

16. CYKLUS JARNÍ ŠKOLY VILÉMA MATHESIA, 19. - 30. března 2001, Praha

V době od 19. do 30. března 2000 se uskutečnil 16. cyklus jarní školy Viléma Mathesia (http://ufal.mff.cuni.cz/vmc/vmc_ls16.html.). Kromě domácích vyučujících (viz níže) přijali pozvání odborníci z USA, Německa, Rakouska, Velké Británie. Celkem se zúčastnilo 53 studentů a výzkumných pracovníků (včetně místních účastníků). Většině cizích účastníků z Běloruska, Bulharska, Estonska, Gruzie, Maďarska, Lotyšska, Moldávie, Polska, Ruska, Rumunska, Slovenska a Jugoslávie byl udělen grant pokrývající ubytování, stravu a kapesné ve výši 2000,- Kč. Platící účastníci přijeli z Finska, Francie, a Ruska. Akce byla zajištěna z prostředků poskytovaných Open Society Institute - HESPa Matematicko-fyzikální fakultou Univerzity Karlovy.

Jan Hajič, **Eva Hajičová**, **Petr Sgall** (všichni CKL) a **Frederick Jelinek** (hostující profesor CKL) vedli kursy s následujícími tituly (po řadě): *The Three Layer Tagging Scenario*, *Topic-Focus Articulation and its Semantic Relevance*, *Tectogrammatics as an Interface Level, Language Modelling for Speech Recognition*

V rámci letní školy se na půdě Pražského lingvistického kroužku konala každoroční speciální Jakobsonovská přednáška s názvem *Prosodic Disambiguation of Syntactic and Semantic Ambiguity In English and Italian*, kterou přednesla Prof. Julia Hirschberg z AT&T Labs, New Jersey, USA.

9TH ELSNET EUROPEAN SUMMER SCHOOL ON LANGUAGE AND SPEECH COMMUNICATION, 16.- 27. července 2001, Praha

Letní škola ELSNET je každoročně pořádána organizací European Network of Excellence in Human Language Technologies (ELSNET). Téma letošní školy bylo „Text and speech corpora“. Během dvou týdnů přednášek vystoupilo 15 učitelů, z toho 14 jich bylo ze zahraničí. Proběhlo celkem 10 sérií přednášek, každá z nich složená z 5 lekcí. **Jan Hajič z CKL MFF UK** vedl sérii přednášek na téma „Linguistic Annotation of a Large Corpus: From Morphology to Syntax“. Školy se zúčastnilo 49 studentů z 23 zemí. Významnou součástí letní školy byla studentská sekce, ve které studenti prezentovali své vědecké zájmy a svou práci.

Přednášky, jejichž součástí byla i praktická cvičení na počítačích, probíhaly v budově MFF UK na Malostranském náměstí paralelně ve dvou posluchárnách a ve dvou počítačových laboratořích. Počítače byly přístupné účastníkům letní školy po celý den.

Na **vedení organizace akce** (zajištění registrace, finanční stránky akce, ubytování studentů a učitelů, přednáškových místností, www stránek, stravování atd.) se od počátku roku 2001 podíleli: **Kiril Ribarov** a **Alena Böhmová** s podporou **Eva Hajičové**. V průběhu akce se **na organizaci podíleli**: **M. Fučík, P.Pajas, D. Zeman** (zajištění provozu laboratoří a poslucháren), **V. Řezníčková, M. Lopatková – Straňáková** (studijní materiály a služby účastníkům), **N. Peterek, J. Štěpánek, J. Havelka** (stravování a kulturní program), L. Brdičková (účetnictví a komunikace s vedením fakulty), **J. Mírovský** (zajištění přestávek). Všichni zúčastnění zaměstnanci CKL se zapojili do organizace zodpovědně a vytvořili účastníkům příjemné zázemí. Organizační zajištění a odborná úroveň školy byla účastníky vysoce hodnocena (viz stať v časopise ELSNET News)

Stránky www letní školy (http://ufal.ms.mff.cuni.cz/~ess2001/) informovaly účastníky a zájemce během přípravy akce o programu, způsobu registrace, kulturních akcích a grantech poskytovaných EU. V průběhu letní školy se nám podařilo zajistit, aby na stránkách bylo možné získat veškeré studijní materiály ke všem přednáškám. Stránky se tak staly užitečným zdrojem informací i po skončení akce.

Akce byla financována z účastnických poplatků a částečně z prostředků grantu IHP EU uděleného organizaci ELSNET.

V nakladatelství Springer se připravuje „tutoriálová“ publikace pokrývající přednášky z letní školy. Za redakci i technickou editaci odpovídá CKL.

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

PŘÍLOHA B *Publikace a přednášky*

UK MFF

Hajič, Jan, Hajičová, Eva, Holub, Martin, Pajas, Petr, Sgall, Petr, Vidová-Hladká, Barbora, Řezníčková, Veronika: The Current Status of the Prague Dependency Treebank. In: TSD2001 Proceedings (eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer), LNAI 2166 Springer-Verlag, Berlin-Heidelberg-New York, pp. 11-20, 2001.

Hajič, Jan, Krbec, Pavel, Květoň, Pavel, Oliva, Karel, Petkevič, Vladimír: Serial Combination of Rules and Statistics: A Case Study in Czech. In: Proceedings of ACL'01, Toulouse, Kontakt, 2001.

Hajič, Jan, Pajas, Petr, Vidová-Hladká, Barbora: The Prague Dependency Treebank: Annotation Structure and Support. In: Proceeding of the IRCS Workshop on Linguistic Databases, Univesity of Pennsylvania, Philadelphia, USA, pp. 105-114, 2001.

Hajič, Jan: Statistické modelování a automatická analýza přirozeného jazyka (morfologie, syntax, překlad). In: Slovenčina a čeština v počítačovom spracovaní (zborník referátov zo seminára Bratislava 26.-27.10.2001 (ed.A. Jarošová)), VEDA, vydavateľstvo SAV, Bratislava, ISBN 80-224-0692-9, 2001.

Hajičová, Eva: Čeština a počítače (Abstrakt). In: Sborník ke konferenci ZNALOSTI 2001, 19-21.6.2001, VŠE, Praha, pp. 307, 2001.

Hajičová, Eva: Syntaktický výzkum nad Českým národním korpusem. In: Čeština - univerzália a specifika 3 (eds. Z. Hladká, P. Karlík), MU Brno, ISBN 80-210-2532-8, pp. 173-181, 2001.

Hajičová, Eva: Information Structure and Syntactic Complexity. In: Proceedings of FDSL 4, Potsdam, (v tisku).

Hajičová, Eva, Havelka, Jiří, Sgall, Petr: Discourse Semantics and the Salience of Referents. Journal of Slavic Linguistics (submitted).

Hajičová, Eva, Panevová, Jarmila, Sgall, Petr: Manuál pro tektogramatické značkování (3. verze, říjen 2001), UFAL Technical Report TR-2001-12, 2001.

Hajičová, Eva, Panevová, Jarmila, Sgall, Petr: Tectogrammatics in corpus tagging. In: Perspectivs on Semantics, Pragmatics, and Discourse, A Festschrift for Ferenc Kiefer (eds I. Kenesei, R. M. Harnish), Pragmatics and Beyond New Series, Vol.90, John Benjamins Publishing Company Amsterdam/Philadelphia, ISBN 90 272 5109 6, pp. 294-299, 2001.

Hajičová, Eva, Sgall, Petr: A reusable corpus needs syntactic annotations: Prague Dependency Treebank, Lancaster, pp.37-48, 2001.

Hajičová, Eva, Sgall, Petr: Dependency, Coordination, and Projectivity. In: Slovo v tekste i v slovare (sbornik statej k semidesjatiletiju akademika Ju.D.Apresjana), (eds. L.L.Iomdin, L.P.Krysin), Studia Philologica. Izd., Jazyki russkoj kul'tury, Moskva. ISBN 5-7859-0199-4, pp.456-466, 2001.

Hajičová, Eva, Sgall, Petr: Topic-focus and salience. In: Proceedings of 39th Annual Meeting of the Association for Computational linguistics, July 9th-11th, CNRS, Toulouse, pp. 268-273, 2001.

Holub, Martin, Míka, Pavel: MATES - An Experimental Linguistic Database System. In: Proceeding of the IRCS Workshop on Linguistic Databases, Univesity of Pennsylvania, Philadelphia, USA.

Kopeček, I., Pala, K., Straňáková-Lopatková, Markéta: Ambiguity Problems in Human-Computer Interaction. In: Proceedings of the conference UAHCI, vol.3 (ed. C. Stephanidis), LEA, Mahwah, New Jersey, ISBN 0-8058-3609-8, pp.486-490, 2001.

Kučerová, Ivona: Teoretická lingvistika a statistické zpracování přirozeného jazyka. In: sborník řady Linguae bohemicae studentinum IV. (v tisku).

Panevová, Jarmila: Některé typy chyb ve stylu odborném a žurnalistickém a možnost jejich automatického odstranění. In: TERMINA 2000, Sborník příspěvků z II. konference 1996 a III. konference 2000, Galén Praha, ISBN 80-7202-105-X, pp. 40-47, 2001.

Panevová, Jarmila: Problémy reflexivního zájmena v češtině. In: Přednášky z XLIV. běhu Letní školy slovanských studií (ed. J. Nehasil), FF UK, Praha, ISBN 80-7308-004-4, pp.81-88, 2001.

Panevová, Jarmila, Řezníčková, Veronika: K možnému pojetí všeobecnosti aktantu. In: Čeština - univerzália a specifika 3 (eds. Z. Hladká, P. Karlík), MU Brno, ISBN 80-210-2532-8, pp. 139-146, 2001.

Sgall, Petr: A remark on Semantics and Pragmatics in Natural Language, PBML 76, MFF UK (v tisku).

Sgall, Petr: Functional Generative Description, Word Order and Focus. Theoretical Linguistics 27, pp.3-19, 2001.

Sgall, Petr: Ohlédnutí pražského lingvisty za dvacátým stoletím. Slovo a slovesnost 62, č. 4, 2001.

Sgall, Petr: Structural and Formal Linguistics in Prague (Preface). In: Towards a Relational - Perspective Approach to Syntactic Semantics, ISBN 7-107-14429-4, pp. xxiii-xxxxviii, 2001.

Sgall, Petr: Volnost jako univerzální vlastnost jazyka. In: Čeština - univerzália a specifika 3 (eds. Z. Hladká, P. Karlík), MU Brno, ISBN 80-210-2532-8, pp. 49-57, 2001.

Skoumalová, Hana, Straňáková-Lopatková, Markéta, Žabokrtský, Zdeněk: Enhancing the Valency Dictionary of Czech Verbs: Tectogrammatical Annotation. In: TSD2001 Proceedings (eds. V. Matoušek, P. Mautner, R. Mouček, K. Taušer), LNAI 2166 Springer-Verlag, Berlin-Heidelberg-New York, ISBN 3-540-42557-8, pp. 142-149, 2001.

Straňáková-Lopatková, Markéta: Homonymie předložkových skupin v češtině a možnost jejího automatického zpracování.ÚFAL Technical Report TR-2001-11, 2001.

Straňáková-Lopatková, Markéta: Ambiguity of Prepositional Groups: Classification, Criteria and Method for Automatic Processing. In: On Prepositions (eds. L. Šaric, D. F. Reindl), Studia Slavica Oldenburgensia 8, Bibliotheks- und Informationssystem, Oldenburg, pp.263-282, 2001.

Straňáková-Lopatková, Markéta: Některé typy syntaktické homonymie (z hlediska možnosti automatického zpracování). In: Čeština - univerzália a specifika 3 (eds. Z. Hladká, P. Karlík), MU Brno, ISBN 80-210-2532-8, pp. 183-195, 2001.

Zeman, Daniel: How Much Will a RE-based Preprocessor Help a Statistical Parser? In: Proceedings of International Workshop on Parsing Technologies, Tsinghua University Press, ISBN 7-302-04925-4, pp.253-256, 2001.

Zeman, Daniel: Parsing with Regular Expressions: A Minute to Learn, a Lifetime to Master. In: PBML 75, MFF UK, Praha, pp.29-37, 2001.

Žabokrtský, Zdeněk: Automatic Functor Assignment in the Prague Dependency Treebank. ÚFAL Technical Report TR-2001-10, 2001.

recenze

Böhmová, Alena, Ribarov, Kiril, Vidová-Hladká, Barbora: D. Biber, S. Conrad and R. Reppen: Corpus Linguistics. Investigating Language Structure and Use. Cambridge Approaches to Linguistics. Cambridge University Press: Cambridge 1998, PBML 76, MFF UK (v tisku).

Havelka, Jiří: K. von Heusinger, U.Egli (eds): Reference and Anaphoric Relations. Studies in Linguistics and Philosophy 72, Kluwer Academic Publishers: Dordrecht, The Netherlands. ISBN 0-7923-6070-2, PBML 75, MFF UK, Praha, pp. 97-100, 2001.

Sgall, Petr: Galina P. Neščimenko: Etničeskij jazyk. Opyt funkcional'noj differenciacii. Specimina philologiae Slavicae, vol.121, 1999. Slovo a slovesnost 62, č.1pp. 71-74, 2001.

Sgall, Petr: H. Filip: Aspect, Eventuality Types and Nominal Reference. Garland Publishing, New York - London 1999. Slovo a slovesnost 62, č.2 pp. 126-130, 2001.

Štěpánek, Jan: CD-ROM Prague Dependency Treebank 1.0., Institute of Formal and Applied Linguistics & Linguistic Data Lab. Published by Linguistic Data Consortium, University of Pennsylvania. PBML 76, MFF UK (v tisku).

ZČU

Psutka, J., Ircing, P., Radová, V.: Experiments with the Recognition of Highly Inflected Spoken Language (Czech) in the Large Vocabulary Task. In: *The 5th World Multiconference on Systemics, Cybernetics* SCI'2001, Orlando, U.S.A., 2001, pp. 559-564.

Ircing, P., Psutka, J.: Two-Pass Recognition of Czech Speech Using Adaptive Vocabulary. In: Text, Speech and Dialogue. *The 4th International Workshop on TSD'2001*, Berlin, Heidelberg, Springer-Verlag. pp.273-277. 2001a.

Ircing, P.: *Language Modeling of Highly Inflectional Language (Czech).* PhD Study Report. Katedra kybernetiky, Centrum komputační lingvistiky, FAV ZČU, Plzeň, 32s. 2001b.

Jelinek, F., Byrne, W., Khudanpur, S., Hladká, B., Ney, H., Och, F.J., Curin, J., Psutka, J.: Robust Knowledge Discovery from Parallel Speech and Text Sources. In: *Proceedings of the Human Language Technology Conference HLT2001*, California, San Diego, 2001.

ÚJČ

Uhlířová, L.: The Case of Czech possessive adjectives and their head nouns: some distributional properties. Glottometrics (www.glottometrics.de), č. 2, 2001, s. 1-9.

Štícha, F.: Kritéria gramatičnosti (Korpus jako argument a inspirace), *Slovo a slovesnost, LXII*, 2001, s. 161-175.

Štícha, F.: Grammar theory and Grammar research. Přednáška na výročním zasedání Societas Linguistica Europaea (SLE), Leuven, srpen 2001.

Štícha, F.: Obligatorní, preferenční a fakultativní užívání subjektového zájmena já po spojkách hypotaktických. Konference Univerzália a specifika, Brno, 2001.