MATHESIOVSKÉ SEMINÁRE
Euro Summer School – Vilém Mathesius Lecture Series 19

Alexander Horák – Martin Ološtiak – Martina Ivanová – Lucia Gianitsová

Správa bude publikovaná v Jazykovednom časopise (roč. 55, 2004, č. 1)

V dňoch 8. – 20. marca 2004 sa v pražskom hoteli Krystal uskutočnilo devätnáste pokračovanie svetovo známeho cyklu prednášok Euro Summer School, tzv. Mathesiovské semináre. Toto podujatie sa konáva pod organizačnou kuratelou Strediska Viléma Mathesia pre výskum a vzdelávanie v oblasti semiotiky a lingvistiky a Ústavu formálnej a aplikovanej lingvistiky Matematicko-fyzikálnej fakulty Univerzity Karlovej v Prahe.

Vysokú odbornú úroveň každoročne garantujú prednášatelia z celého sveta. Tentoraz ponuku prijali: Vladimir Borschev (Rusko), Nicoletta Calzolari, Guglielmo Cinque (Taliansko), Manfred Krifka (Nemecko), Bente Maegaard (Dánsko), Makoto Nagao (Japonsko), Radu Florian, Frederick Jelinek, Aravind Joshi, Barbara Partee, Stuart Shieber (USA), František Čermák, Jan Hajič, Eva Hajičová, Veronika Kolářová, Petr Sgall (Česká republika). Spomínané vedecké autority si pripravili prednášky na rôzne, takmer napospol interdisciplinárne uchopené témy, medzi ktorými dominovala počítačová a korpusová lingvistika, problematika strojového prekladu a pragmatika.

Prednášky prvého týždňa otvoril Makoto Nagao (Kjótska univerzita), ktorý vo svojich troch prednáškach priblížil oblasť strojového prekladu. Najprv zhrnul jeho históriu, kde opisoval najmä systém strojového prekladu založeného na pravidlách (Rule-Based Machine Translation). Tento systém sa považuje za štandardný a jeho základom je mechanizmus pozostávajúci z pravidiel nekontextovej gramatiky a jednoduchého slovníka. V procese prekladu sa pomocou nich uskutočňuje analýza viet východiskového jazyka a následne štrukturálna transformácia, ktorej výsledkom je generovanie syntaktickej štruktúry cieľového jazyka. Napriek pomernej úspešnosti tohto systému sa však objavil aj rad jazykových javov (obzvlášť pri strojovom preklade z japončiny do angličtiny a naopak), ktoré nie je schopný spracovať. Preto bol vyvinutý strojový preklad na princípe analógie, alebo strojový preklad založený na príkladoch (Example-Based Machine Translation), ktorého autorom je práve M. Nagao. Predstavenie tohto systému bolo aj jadrom ďalších prednášok japonského vedca. Oproti doterajším systémom založeným na pravidlách a na prekladovej pamäti, jeho výhodou je zapojenie prekladového frázového lexikónu, jeho koordinácia s prekladovou pamäťou a

použitie dynamického programovania na utváranie analógií v prekladovom frázovom lexikóne. Na záver M. Nagao predstavil aktuálny stav prác v oddelení NLP Kjótskej univerzity, poukázal najmä na metódy extrakcie prekladovej pamäte z paralelných japonsko-anglických korpusov a načrtol aj možné smery ďalšieho výskumu v tejto oblasti.

Najnovšie trendy i vlastné predstavy o využívaní jazykových zdrojov predstavila vo svojich prednáškach Nicoletta Calzolari (Národná akadémia vied, Inštitút komputačnej lingvistiky v Pise). Zdôraznila najmä aspekty vzťahu počítačových lexikónov a textových korpusov, ktoré sa pri pohľade na lexikálny priestor dopĺňajú. „Úplný" počítačový lexikón by mal obsiahnuť všetky naše „poznatky o svete" a mal by ich reprezentovať explicitne a formálne. Calzolari konštatuje, že dosiahnuť úplnosť v rámci „statického" lexikónu je teoreticky nemožné. Preto je potrebné, aby jazyková infraštruktúra obsahovala okrem tradičného lexikónu aj dynamický systém schopný obohatiť slovník o informácie získané on-line z bohatého korpusu, ktorý zachytáva „skutočne realizované" možnosti, množstvo variácií a flexibilitu obsiahnutú v reči. Táto požiadavka je výzvou pre sémantickú anotáciu a autorka ju využíva na predstavenie niekoľkých praktických problémov, ktoré pri sémantickej anotácii anglického a talianskeho textu vznikajú. V závere svojich prednášok N. Calzolari dospela k nevyhnutnosti novej generácie jazykových zdrojov – sémantických sietí, ktoré umožňujú voľný prístup k spoločným poznatkom za účelom lingvistického spracovania.

V prvom týždni sa problematike determinačnej frázy a prenominálneho charakteru vzťažných konštrukcií z generatívno-konfrontačného hľadiska venoval taliansky profesor Guglielmo Cinque z Univerzity v Benátkach. V prednáškach sa zameral na porovnanie prenominálnych a postnominálnych adjektív a prípadnú ambiguitu ich interpretácie v angličtine a taliančine, na problematiku vyjadrovania kategórie určenosti v germánskych a románskych jazykoch a na aplikáciu Greenbergovho pravidla o ľavo-pravej asymetrii prvkov (demonstratív, čísloviek a opisných adjektív), vyskytujúcich sa pri substantívach. Na tomto základe sa pokúsil vymedziť typologické aspekty vzťažných viet – ich štruktúrne umiestnenie v determinačnej fráze, derivačný aspekt prejavujúci sa pri rekonštrukčných procesoch a pod.

Radu Florian (IBM TJ Watson Research Center, Yorktown Heights) v prednáškovom trojbloku nazvanom *Tagging in Natural Language Processing* sa zameral na rôzne spôsoby automatického značkovania pri rozpoznávaní prirodzeného jazyka, ktorého hlavnou úlohou je „pochopenie" ľudskej reči počítačmi. Jestvuje niekoľko aplikácií, prostredníctvom ktorých možno zaznamenať lexikálnosémantické a syntaktické funkcie slov vo vetách. Tieto aplikácie vyžadujú slovnodruhovú charakteristiku daného slova, identifikáciu základných vetných celkov (menná časť a slovesná časť), povrchový parsing, rozpoznanie pomenovaného

denotátu či dezambiguáciu danej lexie viacvýznamových jednotiek. Keďže uvedené informácie týkajúce sa jednotlivých slov vo vete sú navzájom úzko prepojené, tieto aplikácie možno chápať ako súbory sekvenčne zoradených informácií. Pri počítačovom rozpoznávaní prirodzeného jazyka je dôležité, aké sekvenčné modely sa za daných okolností môžu v jazykových prejavoch vyskytnúť. R. Florian predstavil niekoľko metód (lineárne metódy, maximálna entropia, generatívne modely, Markovove pravdepodobnostné modely a i.), prostredníctvom ktorých možno efektívne určiť pravdepodobnosť výskytu danej sekvencie v komplexných jazykových jednotkách, vo vetách. Pravdaže, uvedené metódy vyžadujú implementáciu veľkého množstva vstupných informácií o fungovaní a vzájomnej interakcii lexikálnych jednotiek.

Priekopník generatívno-transformačnej lingvistiky Aravind K. Joshi (Univerzita v Pennsylvánii) sa v troch prednáškach zameral na dve témy. Pri prvej, ktorá mala názov *Starting With Complex Primitives Pays Off: Complicate Locally, Simplify Globally*, najprv uviedol základné princípy generatívno-transformačného prístupu v lingvistike – pojem prvku, prepisovacieho pravidla a gramatiky, ktorá je rekurzívnym systémom generujúcim prepisovacie pravidlá. Následne argumentoval v prospech prístupu, ktorý by zväčšil robustnosť takéhoto systému tým, že by sa komplexnosť, ktorá v generatívnej lingvistike spočíva v komplikovaných pravidlách transformácie, lokalizovala už na úrovni prvkov – komplexných „primitívov“. Tieto „primitívy“ reprezentujú základné – stromové štruktúry, ktoré sú zasa spojením nominálnych, verbálnych, adverbiálnych fráz a pod. Operácie, ktorými sa kombinujú základné štruktúry, sú pripájanie (adjoining), derivácia (derivation), kompozícia (composition). Pri druhej téme, nazvanej *The Boundary Between Sentence and Discourse: Annotation of Discourse Connectives and their arguments*, prešiel od opisu štruktúry vety k opisu štruktúry diskurzu, pričom zostal v teoretickom rámci lexikalizovanej gramatiky spájania stromov.

Zaujímavou oblasťou praktického využitia jazykovedných výskumov je tvorba kontrolovaného jazyka (controlled language – CL). Ide v podstate o verziu jazyka s obmedzenou gramatikou (jednoduchá vetná štruktúra) a so slovnou zásobou zameranou na istú oblasť (často bez polysémie, homonymie či synonymie). Pomocou takéhoto jazyka by mal pisateľ dodržať istú konzistenciu, jednotný štýl a terminológiu v istých typoch textov. Reálne teda ide o súpis pravidiel na vytvorenie textu. Tejto problematike sa vo svojich prvých prednáškach venovala Bente Maegaard z Centra pre jazykovú technológiu (Univerzita v Kodani), ktorá predstavila najmä projekt AECMA. Mnohé obchodné spoločnosti využívajú CL ako prostriedok, ktorý nielenže podporuje, ale často aj usmerňuje tvorbu textov spoločnosti. Nejde len o to, aby texty obsahovali správnu terminológiu. Zároveň je to

možnosť, ako podporiť jednotný spôsob vyjadrovania (tone-of-voice), teda reprezentovať spoločnosť pomocou jazyka. Prednášajúca oboznámila poslucháčov s dánskym projektom VID, ktorý sa sústreďuje na testovanie, rozvoj a na zlepšenie techník zabezpečujúcich vyššiu kvalitu rôznych dokumentov, ktoré produkujú tri veľké škandinávske spoločnosti. Tie používajú automatickú podporu pri písaní návodov na použitie či listov klientom tak, aby dôstojne reprezentovali firmu a jej imidž. Takéto nástroje sú založené práve na technikách CL. Tvorcovia sa snažia vyjsť v ústrety dvom protichodným požiadavkám: 1. vytvoriť nástroj, pomocou ktorého možno vyprodukovať kvalitný (objektívny) text; 2. tento nástroj však musí zabezpečiť vznik textu, ktorý zvýrazňuje subjektívny (jedinečný) profil spoločnosti, teda ju reprezentuje. V poslednej prednáške sa dánska lingvistka venovala populárnemu problému jazykových zdrojov, keďže ide o dôležitý predpoklad rozvoja lingvistických technológií i nevyhnutný materiál pri zdokonaľovaní a testovaní existujúcich nástrojov. Preto je nevyhnutné poznať validitu jazykových zdrojov, ktorá sa zisťuje na základe rozličných kritérií. B. Maegaard predstavila najmä smer, ktorým sa pri overovaní validity jazykových zdrojov uberá European Language Resources Association (Európska asociácia pre výskum jazykových zdrojov).

Českú lingvistiku zastupovali dvaja najvýznamnejší predstavitelia teórie funkčného generatívneho popisu a zároveň hlavní organizátori celého podujatia, Petr Sgall a Eva Hajičová. Petr Sgall (Matematicko-fyzikální fakulta, Univerzita Karlova v Prahe) vo svojom prednáškovom dvojbloku predstavil základné východiská hĺbkovej valenčnej syntaxe, ktorá sa v opozícii k Chomského konštitutívnej teórii vymedzuje na princípe závislosti. V tomto prístupe je vzťah medzi významom a výrazom rozčlenený na postupnosť čiastkových vzťahov medzi prostriedkami nižšej roviny a ich funkciami na bezprostredne vyššej rovine. Na pozadí jakobsonovskej koncepcie príznakovosti sa ukazuje, že jadro systému je možné zachytiť všeobecnými pravidlami (na ich základe sa jednotlivé jazyky odlišujú iba repertoárom atribútov a ich hodnôt, a nie štrukturačným rámcom), zatiaľ čo periféria zahŕňa neprototypické javy, dané diferenciami medzi dialektmi, žánrami, štýlmi a pod. Úlohou popisu je špecifikácia tých pravidiel, ktorými sa zachytáva vzťah medzi hĺbkovou vetnou štruktúrou (tektogramatika) a jej povrchovou, morfematickou reprezentáciou (fenogramatika). Vzťah medzi tektogramatickou rovinou a možnosťami jej povrchovej reprezentácie sa stáva zároveň základom typológie jazykov, ktorej sa P. Sgall venoval bližšie v druhej prednáške.

Interaktívny aspekt jazyka, na ktorý sa vo funkčnom generatívnom popise vzťahuje problematika artikulácie východiska a jadra výpovede, bol predmetom prednášky Evy Hajičovej (Matematicko-fyzikální fakulta, Univerzita Karlova v Prahe). Stanovenie pravidiel na opis informačnej štruktúry vety zahŕňa opozíciu kontextovo voľných

a kontextovo viazaných prvkov vetnej štruktúry vo vzťahu k hĺbkovému slovosledu topiku a fokusu na tektogramatickej rovine vety z aspektu jej syntaktickej štruktúry a indikáciu koreferenčných spojov gramatického a textového charakteru na pozadí stupňov salientnosti z aspektu výpovede ako jednotky diskurzu. Širší teoretický úvod bol východiskom pre využitie artikulácie východiska a jadra výpovede v translatologickej analýze Škvoreckého prózy a jej prekladu do angličtiny.

Problematika teoretickej pragmatiky na pozadí teórie optimality bola predmetom záujmu profesora Manfreda Krifku (Humboldtova univerzita v Berlíne), ktorý začal druhý týždeň prednáškového cyklu. V úvode predstavil klasickú teóriu optimality a jej aplikácie v morfológii. V ďalších častiach sa na pozadí Griceovej teórie konverzačných maxím, neo-griceovských výskumov L. Horna a S. Levinsona a najnovšej práce R. Bluntera a H . Zeevata zaoberal pragmatickou optimalizáciou v komunikácii a na konzekvencie, ktoré tento postup prináša aj vo vzťahu k samotnej teórii optimality. Okrem spomínaných vystúpení M. Krifka na pôde Filozofickej fakulty Univerzity Karlovej svojou prednáškou o sémantike a rečových aktoch prispel do tradičného jakobsonovského cyklu prednášok. Jeho analýza viedla k záveru, že obvyklé oddeľovanie pravdivostne podmienenej sémantiky a teoretickej pragmatiky rečových aktov je problematické, keďže sémantické operátory môžu fungovať v rečových aktoch nielen ako ich deskriptory, ale zároveň aj ako ich performanty.

Oblasti formálnej a lexikálnej sémantiky sa na Mathesiovských seminároch už tradične venujú Barbara Partee (Univerzita v Massachusetts) a Vladimir Borschev (Ruská akadémia vied). V štvordielnom prednáškovom kurze najprv poslucháčov uviedli do problematiky zhrnutím hlavných tém formálnej sémantiky, kde sa pozastavili najmä pri princípe kompozicionality, rozpracovanom ako lingvistami (Katz a Fodor, Jackendoff), tak aj logikmi (Frege, Tarski, Carnap, Montague). Ďalšia prednáška bola zameraná na sémantickú analýzu nominálnych fráz ako všeobecných kvantifikátorov na jazykovom materiáli angličtiny a ruštiny. Zostávajúce dve prednášky boli spojené spoločnou témou – predstavením dialógového systému, ktorý pre svojich študentov anatómie zostavil W. D. Hagamen. Tento systém je počítačový program, ktorý študentom pomáha nielen osvojovať a upevňovať si poznatky z anatómie, ale učí ich aj spôsobu, akým sú tieto poznatky v systéme štruktúrované. Práve na tento aspekt sa prednášatelia zamerali a predstavili svoje rozvinutie Hagamenovho systému do podoby relačnej databázy, v ktorej sa vyskytujú objekty (pojmy ako kosti, svaly, tepny, nervy a pod.) a binárne vzťahy, ktoré ich spájajú (napätie, kontrakcia, x je časťou y a pod.). Na vyhľadávanie v tomto systéme navrhujú predikátovú logiku prvého rádu.

Každoročným prednášateľom na Mathesiovských seminároch je aj Frederick Jelinek (Univerzita Johna Hopkinsa v Baltimore). Oblasťou, ktorej sa výskumne venuje už tri

desaťročia, je štatistické modelovanie jazyka, čo bolo aj témou jeho tohtoročného vystúpenia. Ako sa ukazuje, napriek pomernej jednoduchosti štatistických modelov jazyka v porovnaní s inými, zložitejšími, najmä konekcionistickými (napr. umelé neurónové siete), alebo gramatickými modelmi (napr. modely riadené pravidlami), miera ich úspešnosti je pomerne vysoká. Dokazuje to ich široké uplatnenie, napríklad aj v komerčnej sfére a každodennom živote – rozpoznávanie a syntéza reči, optické rozoznávanie písma, značkovanie textových korpusov lingvistickými informáciami, strojový preklad atď. F. Jelinek sa v prednáškach pokúšal osvetliť teoretické pozadie štatistických n-gramových modelov jazyka spôsobom prístupným pre lingvistov, vďaka čomu sa nemusel znovu pozerať na prázdne tváre (empty faces), na aké bol z týchto kurzov doteraz zvyknutý.

Skôr všeobecne filozofickým ako rýdzo lingvistickým bolo vystúpenie Stuarta Shiebera (Harvardská univerzita v Massachusetts). V prvej prednáške sa zaoberal Turingovým testom, ktorý, ako je známe, sa považuje za základný test umelej inteligencie. V historickom exkurze porovnal názory obhajcov (Dennett) aj odporcov (Gunderson, Davidson, Searle, Block) Turingovho testu ako postačujúcej podmienky pre inteligenciu, pričom sa vrátil až k Descartovi. Turingov test spočíva vo verifikácii otázky, či stroje môžu myslieť a ako to možno posúdiť z pohľadu človeka. Na tento účel sa používa tzv. imitačná hra, kde osoba zadáva otázky v prirodzenom ľudskom jazyku dvom agentom – inej osobe a súčasne aj stroju, pričom nemá možnosť ich pozorovať. Za prisúdenie atribútu inteligencie sa v tomto teste považuje schopnosť verbálnej interakcie, kde by osoba zadávajúca otázky nerozlíšila ľudskú osobu od stroja. Druhá prednáška bola orientovaná trochu inak – na konštrukciu nástrojov na počítačové spracovanie jazyka. Podľa Shiebera je elementom spájajúcim všetky procesy počítačového spracovania prirodzeného jazyka transformácia rozmanitých dátových štruktúr – reťazcov a stromov. Z tohto aspektu sa spracovanie rečového signálu javí ako transdukcia (konverzia/prevod) rečových vzoriek na reťazce foném a následne na slová, morfologické procesy ako transdukcia reťazcov, sémantická interpretácia ako transdukcia syntaktických stromov na stromy sémantických operácií a pod. Vo svetle tejto myšlienky S. Shieber nastolil otázku univerzálneho formalizmu na transdukciu stromových štruktúr a reťazcov prirodzeného jazyka a predstavil niektoré novšie výsledky v tejto oblasti komputačnej lingvistiky a formálnych gramatík.

František Čermák (Filozofická fakulta, Univerzita Karlova v Prahe) v minulosti vo viacerých prácach ukázal, akým spôsobom možno efektívne spojiť klasické lingvistické metódy – v stredoeurópskom priestore reprezentované najmä štrukturalistickým dedičstvom v podobe odkazu Pražského lingvistického krúžku – s najnovšími jazykovednými trendmi, ktoré v súčasnosti stelesňuje najmä počítačová lingvistika. Inak tomu nebolo ani tentoraz. Vo

svojej prednáške *Text Introducers of Proverbs and Other Idioms* sa zameral na spôsoby, akým sa v textoch uvádzajú rôzne druhy frazeologických jednotiek, idiómov. Uvádzacie výrazy (introducers) podrobil komplexnej, funkčno-štruktúrnej, formálno-sémantickej, textologickej, štatistickej, konfrontačnej i pragmatickej analýze, pričom svoje poznatky prezentoval prostredníctvom bohatého dokladového materiálu z databáz Britského národného korpusu a Českého národného korpusu.

Záverečný deň patril Janovi Hajičovi a Veronike Kolářovej z hostiteľského Inštitútu pre formálnu a aplikovanú lingvistiku MFF UK, ktorí predstavili súčasnú podobu Pražského závislostného korpusu (Prague Dependency Treebank – PDT). PDT je systematicky budovaný na základe hierarchického usporiadania troch anotačných úrovní: morfologickej, analytickej a tektogramatickej. Morfologická úroveň zahŕňa priradenie základného gramatického tvaru (lemy) a morfologické značkovanie (tagging). Na analytickej úrovni sa skúma „povrchová" organizácia viet a výpovedí a prostredníctvom tektogramatickej úrovne možno odhaliť „hĺbkové", významové usporiadanie vetných a nadvetných celkov (druhá a tretia anotačná úroveň sa zobrazujú pomocou stromových štruktúr). Práve v nadväznosti na poznatky získané tektogramatickou anotáciou možno efektívnejšie a v interdisciplinárnej kooperácii skúmať také javy, ako napr. kódovanie a dekódovanie informácie, referencia a koreferencia, používanie a fungovanie metaforických významov a i. (porovnaj aj http://quest.ms.mff.cuni.cz/pdt/Corpora/PDT_1.0/Doc/tect.html).

Podľa vyjadrenia „duše" Mathesiovských seminárov, profesorky Hajičovej, ktorá vo svojom záverečnom príhovore poďakovala všetkým členom organizačného tímu (najmä P. Sgallovi, L. Brdičkovej a M. Čmejrekovi), devätnásta kapitola tohto podujatia by mala zostať kapitolou poslednou. Ostáva len dúfať, že situácia sa predsa len zmení, Mathesiovské stretnutia zaokrúhlia svoju tradíciu a na budúci rok sa uskutoční jubilejný, dvadsiaty ročník. Bola by škoda, keby sa toto ojedinelé a v mnohých ohľadoch podnetné podujatie prestalo po toľkých rokoch organizovať.