M A T H E S I O V S K É S E M I N Á R E

Euro Summer School - Vilém Mathesius Lecture Series 18

V dňoch 9. – 22. marca 2003 sa v Prahe konal tradičný, už 18. ročník odborných prednášok a seminárov Vilém Mathesius Lecture Series. Každoročne ich organizuje Spolok Viléma Mathesia pre vedu a výskum v semiotike a lingvistike v spolupráci s Matematicko-fyzikálnou fakultou Univerzity Karlovej v Prahe. Ako už mená prednášateľov napovedajú, prednášky sa sústredili najmä na oblasť jazykovedy a počítačovej lingvistiky.

Na dvojtýždňových seminároch sa zúčastnili doktorandi a mladí vedeckí pracovníci z krajín východnej a strednej Európy – svojich zástupcov tu malo Poľsko, Maďarsko, Rumunsko, Rusko, Ukrajina, Bulharsko, Česká republika a Slovensko. Frekventanti mali možnosť vypočuť si v priebehu dvoch týždňov odborné prednášky a výstupy dvanástich svetoznámych univerzitných profesorov a vedeckých pracovníkov: USA – Barbara Partee, Frederick Jelinek, Mark Steedman, Bonnie Webber, Julia Hirschberg a Martin Kay, Česká republika – Petr Sgall, Eva Hajičová a Ján Hajič, Nemecko – Helmut Schnelle, Rakúsko – Wolfgang Dressler, Maďarsko – Ferenc Kiefer.

Po slávnostnom otvorení tohtoročného cyklu prednášok Evou Hajičovou, riaditeľkou Spolku V. Mathesia, vystúpil ako prvý Helmut S c h n e l l e (Rohr-University, Bochum) s príspevkom *Language in Mind and Brain*, zameraným na tému „jazyk v mysli a mozgu“. Prednášajúci najskôr predložil tézu o možnostiach integrácie lingvistiky s inými vednými odbormi - chémiou, biológiou, fyzikou atď. Pred optimálnou integráciou týchto, svojím zameraním a predmetom výskumu zatiaľ diametrálne odlišných vedných odborov, stojí ešte dlhá cesta. S novými technickými vymoženosťami na zobrazenie mozgu a jeho aktívnych oblastí sa však objavujú stále výkonnejšie nástroje na skúmanie mozgových funkcií pri analýze jazyka. H. Schnelle prezentoval ďalej lingvistiku ako vedu v konfrontácii s výnimočným mozgovým potenciálom zobrazovania a prijímania informácií: zostane lingvistika v izolácii a ponechá výskum mozgu na psychológiu a biológiu, alebo sa začne podieľať na budúcich výskumoch spracovania jazyka v mozgu, spolu s inými vednými odbormi? Kooperácia s neurobiológiou a neuropsychológiou by znamenala jasné pochopenie vzťahu medzi jazykovednými teóriami a pozorovaniami na jednej strane a teóriou a pozorovaním architektúry mozgu a činnosti mozgu na rôznych analytických úrovniach (mozgové systémy, neurónové siete, skupiny neurónov, jednotlivé neuróny) na strane druhej. Prednášajúci naznačil, že štúdium fungovania jazyka v mozgu a mysli môže viesť k radikálnym zmenám v tradičných lingvistických teóriách ako i v pohľade na počítačovú lingvistiku. Kritike podrobil Chomského názor na lingvistiku, ktorý označil v mnohých bodoch za značne zjednodušený, niekedy až úplne nesprávny.

Pod názvom *Integrating Lexical and Compositional Semantic* predstavila Barbara P a r t e e (Univerzita v Massachusetts) jeden z prúdov generatívnej lingvistiky – montaguovskú gramatiku. V 70. rokoch 20. storočia presadzuje Montague, hlavný predstaviteľ intenzionálnej logiky, formálnu analýzu jazyka. Vychádzajúc z princípov pravdivostnej predikátovej logiky, Montague sa pokúša o sémantickú analýzu slovných tried. Montaguova gramatika obsahuje dve zložky – syntaktickú a sémantickú – obe sú navzájom v jednoznačnom vzťahu. Presne definované syntaktické kategórie sú zjednotené tzv. kategoriálnymi pravidlami a operáciami, ktoré konštruujú frázovú gramatiku. Vety, formulované zodpovedajúcimi sémantickými pravidlami, sú ďalej reprezentované na základe princípov mnohých predikátových výpočtov (jedným z najdôležitejších v intenzionálnej logike je tzv. *lambda-calculus*, ktorý umožňuje konverziu či kombináciu viacnásobného predikátu). Tak ako generativisti, pokúšajúci sa formálne vyjadriť a opísať hĺbkovú štruktúru jazyka so všetkými jeho významovými odtienkami, aj Montaguovi prívrženci si v dnešnej

dobe uvedomujú mnohé nedostatky pri pokusoch formálne vyjadriť hovorený jazyk či bežnú ľudskú komunikáciu.

Frederick J e l i n e k v prednáške *Language Modeling for Speech Recognition* oboznámil poslucháčov so základnými teoretickými princípmi systémov rozpoznávania reči, s typickými problémami, ktoré vznikajú pri vývoji takýchto systémov a s metódami, ako sa bežne tieto problémy riešia. Ďalej sa venoval matematickej analýze *n-gramovej* metódy rozpoznávania ambiguít v reálnom čase na základe odhadu pravdepodobnosti výskytu nasledujúceho slova v prichádzajúcom prúde reči. Podrobne vysvetlil úskalia a nedostatky používanej metódy, ktorá je vlastne jediným spôsobom, ako pracujú bežne rozšírené, komerčné systémy rozpoznávania reči.

Prvý deň prednášok uzavrel Petr S g a l l príspevkom *A Dependency-Based Underlying Syntax and the Simple Patterning of the Core of Language* v ktorom priblížil niekoľko známych teórií pražského lingvistického kolektívu. Sústredil sa na centrum a periférium jazykového systému (*core of language* a *periphery of language*), za k centru najbližšiu označil syntax (*underlying syntactic structure*). Prototypickými nástrojmi na vyjadrenie gramatických hodnôt sú morfémy. Pri analýze syntaktickej štruktúry vety nemožno vynechať slovosled a s ním úzko súvisiace východisko a jadro výpovede (*topic and focus of articulation*). Každý jazyk možno rozdeliť do niekoľkých úrovní, jednotlivé úrovne majú vždy svoju vlastnú syntax. Syntax, sémantika a pragmatika teda spolu úzko súvisia a nemožno ich považovať za tri rôzne roviny jazyka. Na príklade sémantického trojuholníka demonštroval rozdiely medzi jazykom ako systémom (*langue*), komunikáciou či diskurzom (*parole*) a kognitívnou zložkou jazyka.

Eva H a j i č o v á sa v prednáške nazvanej *Topic-focus and salience* venovala teórii diskurzu. Príspevok rozdelila na päť častí. Prvé štyri, zamerané skôr teoreticky, mali poslucháčov uviesť do problematiky vetnej stavby: veta sa skladá z dvoch častí – východiska a jadra výpovede (*topic and focus of articulation – TFA*). Presné určenie týchto základných sémantických informácií o vetnej stavbe slúži ďalej na formálny opis jazyka – v tomto prípade pri syntaktickej (tektogramatickej) anotácii Pražského závislostného korpusu. Na príklade z PZK predstavila štandardný postup anotačného nástroja pri zobrazení stromovej štruktúry vety: najvyšším uzlom je predikát, ktorý na seba viaže ďalšie závislé vetné členy. Predikát ako najdôležitejšia, „nová“ informácia rozdeľuje vetu na *topic* (pre príjemcu už známa informácia, v ľavej časti stromovej štruktúry) a *focus* (nová informácia, v pravej časti stromovej štruktúry). Aby nezostalo len pri teórii, na záver prednášajúca predviedla i praktickú ukážku analýzy diskurzu na príklade literárneho diela Josefa Škvoreckého *Scherzo capriccioso. Veselý sen o Dvořákovi* v porovnaní s jeho anglickou verziou *Dvořák in Love*.[1] V grafe sa najskôr vyznačia vodorovne čísla jednotlivých viet, jeden riadok v grafe = jedna románová veta. Zvisle sa postupne zaznačí každá postava, ktorá sa v danej vete objaví, pričom sa začína vždy zľava. Ďalej sa podľa výskytu v diskurze vedú pre konkrétne postavy akési priamky aktívnosti – čím neskôr má daná postava ďalší prehovor (je aktívna), tým viac sa vzďaľuje jej priamka aktívnosti smerom do prava. Takýmto spôsobom možno získať podrobný rozbor celého diela. Veľmi efektívne využitie sa ukazuje aj pri porovnávaní paralelných (viacjazyčných) textov.

Problematiku komparatívnej morfológie nastolil vo svojich trojdňových vstupoch Wolfgang D r e s s l e r (Jazykovedný inštitút vo Viedni). Pod názvom *Static and Dynamic morphology* predstavil hneď niekoľko naliehavých otázok v morfologických výskumoch. Podrobnou analýzou a porovnaním špecifických odlišností medzi slovanskými, germánskymi a románskymi jazykmi demonštroval rozdiely medzi statickou a dynamickou morfológiou, diferencoval pravidelnosť a nepravidelnosť v slovotvorbe, rozlíšil produktívne a neproduktívne zložky v morfológii. Veľmi zaujímavé boli aj názorné príklady fenoménu interferencie pri preberaní či požičiavaní si cudzích slov medzi jednotlivými jazykmi. Možno

[1] podrobnejšie informácie o J. Škvoreckom možno získať na http://www.writersunion.ca/s/skvoreky.htm

konštatovať, že neexistuje stabilné a všeobecne platné pravidlo, ktoré by určilo ďalšie správanie sa konkrétneho neologizmu v cieľovom jazyku.

Úvodný týždeň prednášok uzavrel Jan H a j i č referátom o morfologickej, povrchovo-syntaktickej a tektogramatickej anotácii Českého národného korpusu. V úvode prednášky opísal systém morfologickej anotácie českého jazyka vyvinutý na UK Praha a používaný v Českom národnom korpuse. Nadväzne sa venoval povrchovo-syntaktickej a tektogramatickej anotácii, so stručným objasnením princípov anotácie a postupu, ako anotácia prebieha, ako aj naplánovaných cieľov pri tvorbe tektogramaticky anotovaného korpusu, vhodného na trénovanie NLP[2] programov založených na štatistických princípoch.

Mark S t e e d m a n (Fakulta informatiky Edinburskej univerzity) začal druhý týždeň Mathesiovských seminárov príspevkom *Coordination and the Theory of Grammar*. Prezentoval niekoľko aspektov spracovania prirodzeného jazyka, v prvej časti nazvanej Koordinácia a teória gramatiky poslucháčov stručne uviedol do problematiky spracovania prirodzeného jazyka. Na viacerých príkladoch objasnil teóriu kombinačnej kategoriálnej gramatiky a prezentoval jej formalizmus. V druhej časti sa venoval formalizovaniu intonácie a vysvetlil výhody kombinačnej kategoriálnej gramatiky na zachytenie intonačnej a informačnej štruktúry jazyka. Trojdielnu sériu prednášok ukončil teóriou o rozpoznávaní a spracovávaní prirodzeného jazyka metódou kombinačnej kategoriálnej gramatiky. Opísal architektúru takéhoto parsera[3], vymedzil jeho problémy a definoval pravdepodobnostné modely kategoriálnej kombinačnej gramatiky. Na záver prezentoval konkrétne hodnoty úspešnosti týchto modelov.

Bonnie W e b b e r (Fakulta informatiky Edinburskej univerzity) sa v prenáške *Computing Discourse Structure and Discourse Semantics* opäť vrátila k teórii diskurzu a reláciám medzi jeho zložkami. Cieľom výskumov skupiny vedcov z Edinburskej univerzity je porozumieť, v akom rozsahu môže slúžiť anafora ako mechanizmus na vyjadrenie významu v reči; ako takéto anaforické vyjadrenie ďalej interaguje s pokrytím významu reči cez kompozičnú sémantiku a inferenciu. Veľká časť prednášky bola venovaná príslovkám ako spájacím výrazom vo vetnej štruktúre a ich sémantickému významu. Predstavila i *lambda-formalizmus* ako jeden z vhodných nástrojov na formálne vyjadrenie týchto výrazových prostriedkov. Podrobne tiež opísala mechanizmus pre následné včlenenie tohto formálneho vyjadrenia do automatického počítačového spracovania jazyka.

Prvá prednáška Ferenca K i e f e r a (Maďarsko) *Event Structure and Aspect* sa konala v rámci Jacobsonových prednášok, ktoré sú pravidelne organizované v priestoroch Filozofickej fakulty UK. Ďalšie dve časti, nazvané *Three Ways of Doing Semantics*, prebiehali už v rámci Mathesiovských dní. Prednášajúci pútavým a zaujímavým spôsobom rozobral prístupy skúmania sémantických vlastností objektov v rámci troch hlavných prístupov: Prvý prístup sa zameriava na informačnú hodnotu výpovede, kde zmysel leží hlavne vo vzťahoch medzi symbolmi a objektami, ktoré reprezentujú. Druhý prístup sa zameriava na kognitívny význam jazyka, kde zmysel symbolov spočíva v ich internej reprezentácii v mysli človeka. Tretí prístup zdôrazňuje význam komunikácie ako sociálnej aktivity a vzájomnej interakcie.

Neplánovite, namiesto jednej prednášky Nicoletty Calzolari, ktorá sa na seminári nemohla zúčasniť, vystúpil Michail B o l d a s o v (Štátna Lomonosovova univerzita, Moskva) s referátom *Multilingual Generation in Data Base NL-Interface*. Prítomným popísal komerčný systém InBASE[4] – je to systém spájajúci SQL databázu s rozhraním, kde je možné zadávať queries[5] v prirodzenom jazyku. V súčasnosti je systém InBASE použiteľný v troch jazykoch, a

[2] Natural Language Processing

[3] nástroj pre automatickú syntaktickú analýzu

[4] http://www.inbase.artint.ru

[5] *query* je otázka na databázový systém, obvykle položená prostredníctvom na to určeného špeciálneho formálneho počítačového jazyka (ale špecificky **nie** v tomto prípade) a rozhrania

to v ruštine, angličtine a nemčine. Súčasťou systému je aj NLG[6] modul, ktorý refrázuje zadanú otázku v prirodzenom jazyku, čo môže slúžiť ako efektívna kontrola „či nás počítač správne pochopil“. Keďže prednáška bola pripravená neplánovane, poslucháči s pochopením vzali na vedomie fakt, že na dostupnom počítači nebolo možné predviesť fungujúci model v ruštine. M. Boldasov prezentoval prototyp na angličtine – pravidlá pre generovanie jazyka napísal, podľa vlastných slov, „počas jednej hodiny včera v noci“. O to viac zapôsobila kvalita a správnosť generovaného výstupu.

Martin Kay (Stanfordská univerzita) prezentoval v prednáške nazvanej *Triangulation: An Approach to Partially Automated Machine Translation* problematiku strojového prekladu. Na úvod zdôraznil, že tradičný pohľad na automatický preklad, ktorý považuje skôr za lingvistický problém, je dnes už neobhájiteľný. Správny výber medzi rôznymi alternatívami, ktoré ponúka automatický prekladací systém, vyžaduje určitú dávku všeobecných vedomostí a ľudského citu. Existujú dva prístupy k automatickému prekladu: komerčný a akademický. Kay sa ďalej zameral na akademický prístup. Predstavil typickú štruktúru väčšiny systémov na strojový preklad, obsahujúce morfologické, syntaktické, lexikálne, (niekedy i sémantické) zložky pre pôvodný a pre cieľový jazyk. Jednotlivé vrstvy zabezpečujú spracovanie základných slov, zložených výrazov, viet i významových celkov. Spracúvaná informácia prechádza pri analýze postupne jednotlivými vrstvami – od jednoduchej (morfologickej) až po abstraktnú (sémantickú). Systém ďalej obsahuje časť zodpovednú za prenos; je to jediná časť, ktorá je špecializovaná pre príslušný pár jazykov. Táto časť zabezpečuje konverziu najabstraktnejšej úrovne prekladaného jazyka do zodpovedajúcej úrovne cieľového jazyka. Preložená veta sa získa z tejto úrovne postupom, ktorý je v podstate reverzný k analýze. Napokon Kay prezentoval typickú trasu prekladaných dát, predpokladajúc prítomnosť ľudského korektora a prebral niektoré možnosti optimalizácie tejto trasy.

Posledný blok prednášok na tohtoročných Mathesiovských dňoch patril Julii H i r s c h b e r g (Columbia University and AT&T Labs), ktorá sa venovala systémom rozpoznávania reči. Špecializuje sa na dialógové systémy komunikácie, aké sú bežne používané v USA na poskytovanie informácií zákazníkom. Popri klasickom úvode do *speech recognition*[7] systémov, prehľade najčastejších nedostatkov používaných systémov a perspektív rozvoja bola prednáška venovaná aj trochu zanedbávaným oblastiam rozpoznávania reči, ako je napríklad použitie intonácie a prozódie na získanie dodatočných informácií a spätnej väzby na overenie správnosti predchádzajúceho rozpoznania.

18. ročník Mathesiovských seminárov uzatvorila opäť Eva Hajičová. Vyjadrila poďakovanie všetkým prednášajúcim, aktívnym účastníkom a poslucháčom. Nezabudla tiež vysloviť vďaku spolupracovníkom z Centra Viléma Mathesia.: Petrovi Sgallovi, predsedovi vedeckej rady, Libuši Brdičkovej, koordinátorke, a Martinovi Čmejrekovi, odbornému poradcovi. Aj vďaka nim prebehol ostatný ročník tohto významného lingvistického podujatia na vysokej úrovni.

Pavol DOMIN – Martina FORRÓOVÁ – Radovan GARABÍK

[6]Natural Language Generation
[7]t.j. Systémy automatického strojového rozpoznávania reči