Matematická lingvistika

Ve druhé polovině 20. stol. dochází ve vědě k výrazné změně charakterizované vznikem nových „pomezních“ oborů. Tradiční odvětví vědy přijímají nové metody z oborů blízkých i relativně vzdálených. Objevují se disciplíny, které kombinují otázky lingvistické s tématy a metodami oborů, jako je matematika, psychologie, sociologie, antropologie, ale posléze i informatika nebo umělá inteligence. Vznikají tak disciplíny stojící na hranicích různých oborů, nikoli ovšem na jejich periferii.

Matematická lingvistika vykrystalizovala jako vědní obor na pomezí lingvistiky, matematiky, posléze i umělé inteligence (artificial intelligence – AI) a informatiky. Snaží se o exaktní popis přirozeného jazyka opřený o matematické metody. Z určitých aspektů se dá dále rozčlenit na **komputační (počítačovou) lingvistiku**, zabývající se zkoumáním a zpracováním přirozeného jazyka prostřednictvím počítačů a metod informatiky, **formální lingvistiku**, zabývající se formálním popisem gramatik a jazyků, a **kvantitativní (statistickou) lingvistiku**, využívající statistické stochastické a pravděpodobnostní metody aplikované na přirozený jazyk (dnes rozsáhlé využití pro anotace jazykových korpusů).

V současnosti se ovšem prosazuje termín **počítačové zpracování přirozeného jazyka** (*natural language processing* – NLP), který zahrnuje to, co je uvedeno pod komputační lingvistikou. Celkem samostatně se mluví o tzv. **language engineering** (doslovný překlad tohoto termínu *jazykové inženýrství* je poněkud zatížen, neboť se takto původně v 50. letech 20. století označovala nežádoucí manipulace jazyka při různých pokusech o násilnou kodifikaci), jež pokrývá algoritmické techniky v popisu přirozeného jazyka a softwarové nástroje vznikající jejich implementací, tedy aplikace směřující ke strojovému překladu, automatickému ukládání a vyhledávání informací, tvorbě dotazovacích systémů, gramatických korektorů atp.

V souvislosti s technickým vývojem počítačů dochází od počátku 90. let 20. stol. k prudkému rozvoji korpusové lingvistiky (srov. níže a rovněž kapitolu Korpusová lingvistika).

Vysoká míra homonymie a synonymie výrazů a významová vágnost projevující se na všech jazykových rovinách klade explicitnímu (algoritmickému potažmo strojovému) popisu přirozeného jazyka do cesty řadu překážek. Budování nástrojů pro automatické odstraňování víceznačných interpretací jednotek přirozeného jazyka (*automatická disambiguace*) je tudíž jedním z prvořadých úkolů matematické lingvistiky, jehož vyřešení má dalekosáhlý význam pro rozvoj veškerých aplikací v oblasti NLP, v korpusové lingvistice, strojovém překladu,

systémech směřujících k automatickému porozumění textu v libovolném neomezeném přirozeném jazyce (vyhledávání informací), v automatické analýze a syntéze řeči atd.

Ačkoliv se vznik matematické lingvistiky datuje až od 2. poloviny 20. století, můžeme říci, že některá odvětví mají své prvopočátky již v dřívějším vývoji lingvistické vědy. Počátky matematického přístupu k jazyku se v Evropě vynořují s nástupem tzv. aristotelovského racionalismu. Byl vyvolán překladem Aristotelova díla *Organon* do latiny, k němuž dal podnět Tomáš Akvinský. V jazykovědě vedlo uplatnění racionálně logických postupů ke vzniku prvních „spekulativních" „filozofických" gramatik, jako byla třeba *Summa grammatica* Rogera Bacona usilující postihnout „jazyk vůbec", vlastně jazyk univerzální. Pro Baconova současníka Raimunda Lulla, kolísajícího v matematizujících představách mezi kabalou a logikou, byla univerzálnost jazyka paralelní s univerzálností matematické kombinatoriky, neboť na ní je podle Lulla založen jeho výrazový plán. On sám vytvářel pojmové a písmenkové tabulky a kotouče (pro kombinaci a hierarchizaci pojmů a písemných záznamů), podobně jako Giordano Bruno aj. S tímto pracovním instrumentářem se pracovalo ještě i později, především jeho aplikací na popisy, výklady a praktické využití fakt jazykových, ale i grafických, a to ani ne tak v steganografii (tj. zkoumání tajných písem), jako spíše při vytváření šifrovaných systémů, vlastně svého druhu předchůdců dnešních těsnopisných soustav. Kombinatoriky v pojmově jazykových tabulkách, klasifikačních stromech a kotoučích se užívalo až do raného novověku; nejvýraznějším reprezentantem tohoto postupu byl pravděpodobně v 17. stol. John Wilkins a na přelomu 17. a 18. stol. G. W. Leibniz. Ve filozofických koncepcích jazyka se postupně rezignovalo na vytvoření seznamu ideálních obsahů a jejich kombinací a vztahů a pozornost se obracela stále zřetelněji k logicko-matematickému kalkulu (vlastně tedy k formální syntaxi). Tento obrat lze ve filozofii sledovat od Marie Jeana Antoina Nicolase Caritata, markýze Condorceta až po Ludwiga Wittgensteina. Na české půdě vstoupil do tohoto myšlenkového proudu nejvýrazněji už v 17. stol. logik, matematik a lingvista **Jan Caramuel z Lobkovic**, autor několika pokusů o dokonalý filozofický jazyk. Pro logickou analýzu přirozeného jazyka mělo zásadní význam dílo matematika Gottloba Fregeho.

Na možnosti používání matematických metod v lingvistice upozorňoval již v 1. polovině 19. stol. např. ruský matematik V. J. Buňakovskij a počátkem 20. stol. Jan Baudouin de Courtenay. Z jazykovědců se jako první zabýval kvantitativními jevy v jazyce americký lingvista W. D. Whitney, který zkoumal frekvenci anglických hlásek. Autorem prvního frekvenčního slovníku je německý stenograf F. W. Käding (*Häufigkeits Wörterbuch der Deutschen Sprache*, 1898).

Kvantitativní lingvistiku v 1. polovině 20. stol. ovlivnili zejména ruský matematik A. A. Markov a americký lingvista německého původu George K. Zipf. Andrej Andrejevič Markov vydal v roce 1913 statistickou analýzu textu Evžena Oněgina (Markov, 1913). Toto dílo probudilo zájem o mezioborovou spolupráci lingvistiky a matematiky. Na základě statistického zkoumání výskytu ruských hlásek a pravděpodobnosti, s jakou po sobě následují v textu, došel k závěru, že je možné předvídat pravděpodobnost jejich výskytu („markovův proces"). Šlo o první důslednou aplikaci

matematické statistiky v jazykovědě, na niž pak navázaly další aplikace teorie pravděpodobnosti a teorie informace.

O kvantitativní lingvistice se někdy mluví také jako o Zipfově lingvistice podle amerického lingvisty německého původu George Kingsleyho Zipfa, z Harvardovy univerzity. Zipf zkoumal ve 20. a 30. letech 20. stol. relativní frekvenci hlásek, zajímal se o psychologické a fyziologické faktory ovlivňující produkci a percepci řeči. Upozornil na vztah mezi frekvencí slov a jejich pořadím, dále na vztah mezi frekvencí slova a počtem různých slov, která tuto frekvenci mají, a na vztah mezi frekvencí slova a počtem jeho významů (Zipfovy zákony). Jeho přínos jazykovědě byl objeven s určitým zpožděním. Teprve poté, co matematikové, fyzikové a další přírodovědci odkryli ve svých oborech analogie k zákonům, které Zipf formuloval, prorazily jeho myšlenky také v lingvistice; srov. např. Zipf (1949 aj.), Uhlířová (2002), Hřebíček (2002)

Z výše zmíněných tří hlavních „proudů“ (počítačová, formální, kvantitativní) shrnovaných pod střešní název matematická lingvistika má kvantitativní (statistická) lingvistika o trochu delší tradici. Na české půdě měla své předchůdce v nahodilých, ad hoc pořizovaných statistických pozorováních a závěrech, oblíbených zejména ve filologických, na excerpci dokladů založených pracích pozitivistického zaměření. Koncepčněji aplikoval tuto metodu **Martin Hattala** v sérii studií, v nichž se pokoušel najít jisté pravidelnosti a zákonitosti v hláskové stavbě slovanských slov, zejména v jejich konsonantické kombinatorice v násloví, a to zcela programově s pomocí matematicko-statistických postupů. Ty vedly k explikacím dosti už systematickým a svým způsobem exaktním, ale nikoli bez nebezpečí jistého mechanického přecenění kvantitativních údajů a výsledků statistického vyhodnocení zkoumaného materiálu. Zmiňujeme-li počátky kvantitativní lingvistiky postavené na pevných metodologických základech, nelze opominout práci **Viléma Mathesia** (1911), sledující potencionálnost jazykových jevů. Vychází z toho, že jazyk se neřídí „absolutními zákony“, ale že v řeči každého jednotlivce existuje kolísání „v určitých mezích a s určitou tendencí“. Jazyk jako systém funkčně diferencovaných prostředků existující ve vědomí jednotlivých mluvčích je systém pravděpodobnostní povahy. Zajímavostí mohou být i případy interdisciplinárních úvah o využití metod matematické statistiky ve stylistice a literární teorii (Wolf, 1928). Jako samostatná lingvistická disciplína s vlastním předmětem a metodami se kvantitativní lingvistika začala u nás rozvíjet v rámci strukturalismu a její počátky jsou spojeny s pražskou školou, s pojmem kvanta a funkčního zatížení, srov. J. Vachek (1957), J. Krámský (1959), B. Trnka (1937, 1951), M. Těšitelová (1951 aj.).

Relativním mezníkem pro rozvoj matematické lingvistiky ve světě bylo vydání Chomského spisu o generativní teorii (*Syntactic structures*, 1957), česky vyšla v polovině 60. let. Model jazyka založený na matematických metodách, jímž byla původní Chomského generativní gramatika, má dnes mnoho

variant. O zahrnutí sémantické roviny při formálním popisu přirozeného jazyka se na počátku zasloužili Jerry Alan Fodor – Jerrold Katz (1963), o zahrnutí roviny pragmatické se poprvé zmiňuje Yehoshua Bar-Hillel (1971). K vzájemnému vztahu syntaxe, sémantiky a pragmatiky srov. též Pavel Materna – Karel Pala – Aleš Svoboda (1976, 1979); viz též Mluvnictví.

O aplikaci matematických metod v jazykovědě se zasloužila řada lingvistů bývalého SSSR (O. S. Achmanova, N. D. Andrejev, R. M. Frumkina, I. A. Mel´čuk, J. V. Padučeva aj.). Z východního bloku významně přispěl k rozvinutí teorie analytických modelů rumunský lingvista a matematik Solomon Marcus. Ti měli vliv na rozvoj oboru matematické lingvistiky u nás v 60. letech.

Počítačové zpracování přirozeného jazyka (Natural Language Processing, NLP) se velmi dlouho zaměřovalo především na strojový překlad – Machine Translation – MT (do poloviny 80. let). Od 60. let se badatelé pokouší zachytit realitu jazyka systematicky v gramaticko-logických modelech aplikovaných při strojovém zpracování přirozeného jazyka. Současnost bere gramaticko-logickou kostru jako základ, rozeznává důležitost aktuálního, zvykového a metaforického použití jazyka. Dnes se aplikace NLP orientují nejen na strojový překlad, ale zejména na tvorbu tzv. korpusových nástrojů (manažerů, automatických anotačních programů) a na programy používané v oblasti korpusové počítačové lexikografie.

K probuzení zájmu o spojení matematických metod a výzkum přirozeného jazyka, který promptně reagoval na světový trend, došlo u nás v 60. letech. Překladový sborník *Teorie informace a jazykověda* (1964) byl impulsem pro vývoj oboru a měl vliv na řadu badatelů.

Na Chomského teorii u nás poměrně velmi rychle zareagovala skupina badatelů kolem Petra Sgalla, Pavla Nováka, Dany Konečné a Bohumila Palka, kteří na FF UK v Praze koncem 50. let otevřeli Oddělení teorie strojového překladu. Jejich tehdejší studenti vyrostli v uznávané odborníky na matematickou lingvistiku (Eva Hajičová, Dana Konečná, Ladislav Nebeský, Karel Pala, Jarmila Panevová, Petr Piťha a další). V roce 1964 vychází kniha autorského kolektivu pod vedením **Petra Sgalla** *Cesty moderní jazykovědy* s podtitulem *Jazykověda a automatizace.* Jednotlivé oddíly se zaměřují na jazyk a techniku (strojový překlad, automatické ukládání a vyhledávání informací), algebraickou lingvistiku (formální studium a popis jazyka, generativní gramatiku, rekognoskativní gramatiku, analytické modely jazyka a modely jazykového vývoje a jazykové různosti), kvantitativní lingvistiku (kvantitativní vztahy v lexiku, význam teorie informace a kvantitativních metod pro lingvistiku), mechanizaci a automatizaci (využití počítačů) v lingvistice.

Počátkem 60. let na FF UK v Praze vzniká pojmový rámec pro formální popis přirozeného jazyka označovaný v odborné literatuře termínem *funkční generativní popis* – FGP (anglicky FGD – Functional Generative Description), který je podnes rozvíjen, doplňován a obohacován. Je výsledkem spolupráce celé řady českých badatelů (Eva Benešová-Buráňová,

Eva Hajičová, Květoslava Králíková, Ladislav Nebeský, Pavel Novák, Jarmila Panevová, Petr Piťha, Petr Sgall aj.), plodně čerpá z domácí lingvistické tradice, především z pojetí závislostní syntaxe, jak je nacházíme v pracích představitelů pražské lingvistické školy, ale i z pojetí Vladimíra Šmilauera, navazujícího na Tesnièrovy zásady. Kriticky vstřebává podněty, které rozvinuli ve svých pracích Vilém Mathesius, Vladimír Skalička, Miloš Dokulil, Zdeněk Hlavsa, František Daneš a další. Popis syntaktické roviny jazyka doplňuje o formálně pojatou a rozpracovanou teorii aktuálního členění věty (srov. Sgall – Hajičová – Buráňová, 1980). Tato tradice nezůstává uzavřena sama v sobě a rozvíjí se ovšem na pozadí světového lingvistického dění. Vychází z některých metod generativní syntaxe N. Chomského. Inspiruje se bádáním na poli sémantiky, jak je rozvinuli především R. Montague a B. H. Partee(ová). Interpretace jazyka je ve FGP založena na integrovaném popisu syntaxe, sémantiky a pragmatiky (podrobněji srov. Sgall, 2003). Zásadní význam v 60. letech měla monografie **Petra Sgalla** (1967) *Generativní popis jazyka a česká deklinace*. Jednotliví badatelé uveřejňovali odborné studie např. v *The Prague Bulletin of Mathematical Linguistics*, *Slově a slovesnosti*, *Kybernetice*, *Čs. informatice* a dalších (*Aplikace matematiky*, *Metodika a technika informací*).

Matematická lingvistika je původně spojena s experimenty v oblasti strojového překladu. První pokusy uskutečněné v USA r. 1954 a o rok později v SSSR byly omezeny na doslovný překlad „slovo za slovo“. Nicméně i ony otevřely cestu k uplatnění strojů tam, kde dosud vládl a asi bude i nadále vládnout člověk. Prvotní optimismus vystřídalo vystřízlivění, které však nevedlo k rezignaci, ale spíše k trpělivému pokračování cestou dílčích úspěchů.

Jedním z průkopníků strojového překladu byl i americký lingvista Paul Garvin (studoval v Karlových Varech a Praze), jenž se podílel na pokusu o strojový překlad na Georgetown University ve Washingtonu. Na Harvardu pod vedením A. G. Oettingera a S. Kuna byla experimentálně ověřena tzv. prediktivní (syntaktická) analýza. Z evropských lingvistů uveďme alespoň skupinu K. Brockhause (univerzity Münster, Kostnice, Heidelberg). První prakticky fungující systémy se objevily v Kanadě (METEO – překlady meteorologických předpovědí, projekt TAUM – Traduction automatique à l'Université de Montréal), na ně pak navázal tým na universitě v Grenoblu (skupina GETA – Groupe d'études pour la traduction automatique v čele s Bernardem Vauquoisem). Pro účely strojového překladu se v bývalém SSSR zabývali I. I. Revzin a V. J. Rozencvejg analýzou angličtiny, skupina kolem Ju. D. Apresjana analýzou ruštiny, tým vedený O. S. Kulaginovou problémy strojového překladu z ruštiny do francouzštiny. Z dalších autorů jmenujme alespoň N. D. Andrejeva, V. V. Ivanova, I. A. Mel´čuka, D. J. Panova, S. K. Šaumjana, V. A. Uspenského. A. K. Žolkovského.

Strojovému překladu se dnes věnují především velké počítačové firmy (např. IBM, SIEMENS) a samostatně firma Systran, jež vytvořila překladový systém používaný jako oficiální v EU, i univerzitní týmy v USA, Evropě, Japonsku atd. Význam pro počítačové zpracování menších evropských jazyků pro účely strojového překladu měl velký projekt Eurotran.

Počátkem 90. let byla založena Mezinárodní asociace pro strojový překlad (IAMT) se třemi samostatnými regionálními organizacemi: evropskou (European Association for Machine Translation – EAMT), americkou (Association for Machine Translation in the Americas – AMTA) a asijsko-tichomořskou (Asian-Pacific Association for Machine Translation – AAMT). Tyto organizace sdružují výzkumné ústavy, obchodní společnosti, vědecké pracovníky, odborníky z příbuzných oborů a překladatele, jejichž společným zájmem je strojový překlad. IAMT vydává svá vlastní periodika – Machine Translation, organizuje odborné semináře a jednou za dva roky pořádá mezinárodní konferenci MT Summit.

U nás se první pokus o strojový překlad z angličtiny do češtiny konal v Praze v lednu r. 1960. Skupina lingvistů Karlovy univerzity (Eva Hajičová, Zdeněk Kirschner, Jarmila Panevová, Petr Piťha, Petr Sgall) provedla ve spolupráci s Výzkumným ústavem matematických strojů (VÚMS) tento experiment na počítači SAPO české výroby. Na ně pak navázala řada projektů zaměřených na automatický překlad. V 70. letech projekt APAČ (1977-1986) jehož cílem byla automatizace překladu mezi češtinou a angličtinou, v 80. letech se projekt RUSLAN (1987-1990) zaměřil na strojový překlad z češtiny do ruštiny. Tyto experimenty pokračovaly počátkem 90. let v rámci projektu česko-anglického strojového překladu MATRACE (1990-1992). S úspěchy a úskalími strojového překladu seznamuje čtenáře přístupnou formou kniha *Učíme stroje česky* (1986) **Petra Sgalla – Evy Hajičové – Petra Piťhy**; srov. taky např. Panevová – Sgall (1980/81), Hajičová – Kirschner – Sgall (1981). Publikace *Cesty moderní jazykovědy* (1964) a právě zmíněná *Učíme stroje česky* (1986) byly přínosné nejen pro rozvoj oborů, které shrnujeme pod střešní název matematická lingvistika, ale přispěly rovněž k popularizaci moderních lingvistických trendů.

Po určité přetržce způsobené nepřízní politického vývoje dochází v 60. letech k uvolnění výzkumu na poli kvantitativní lingvistiky, v jehož rámci se vzájemně doplňují a obohacují různé pohledy podmíněné generačně i koncepčně; srov. Těšitelová (1999). Překladový sborník *Teorie informace a jazykověda* (1964) přinesl impulsy pro bádání zaměřené na výzkum kvantitativních vztahů ve slovní zásobě, zpřístupnil novinky referující o významu teorie informace a kvantitativních metod pro lingvistiku a inspiroval teorii i praxi kvantitativní lingvistiky. Kvantitativním metodám užívaným v lingvistickém výzkumu i dalším problémům kvantitativní lingvistiky je věnována kapitola *Kvantitativní lingvistika* v knize *Cesty moderní jazykovědy* (1964).

V roce 1961 vychází první *Frekvenční slovník češtiny* (FSČ), který zpracovali **Jaroslav Jelínek – Josef V. Bečka – Marie Těšitelová**. Na půdě Ústavu pro jazyk český vzniklo dnes už neexistující Oddělení kvantitativní lingvistiky, u jehož zrodu stála Marie Těšitelová a které později vedl Lubomír Doležel. V průběhu 60. – 80. let rozvíjelo toto pracoviště výzkum kvantitativních charakteristik současné češtiny v její psané i mluvené podobě. 70. léta

znamenají u nás určitý průlom v technických možnostech. Vzniká první počítačově čitelný korpus, z nějž vzešla skupina frekvenčních slovníků (viz níže). Jeho autorkou je Marie Těšitelová (ÚJČ ČSAV).

Z prvního elektronického korpusu u nás, zpracovaného pomocí děrných štítků a čítajícího 540 tisíc slovních výskytů, byla čerpána data fonologická, grafématická, morfologická, slovnědruhová, lexikální a syntaktická. Ve sborníku *The Prague Studies in Mathematical Linguistics* (PSML), v odborných časopisech (*Slovo a slovesnost* aj.) a knižních sériích (např. *Glottometrika*, *Quantitative Linguistics*, u jejichž zrodu stál slovenský lingvista Gabriel Altmann) byly průběžně publikovány články řady autorů (Helena Confortiová, Lubomír Doležel, Jan Králík, Jiří Kraus, Marie Ludvíková, Iva Nebeská, Eleonora Slavíčková, Jitka Štindlová, Marie Těšitelová, Ludmila Uhlířová aj.) založené na kvantitativním výzkumu jazyka. Pod vedením **Marie Těšitelové** vyšla řada odborných publikací – *Otázky lexikální statistiky* (1974), *O využití statistických metod v gramatice* (1980), *Kvantitativní charakteristiky současné české publicistiky* (1982), *Kvantitativní charakteristiky současné odborné češtiny* (1983), *Kvantitativní charakteristiky současné češtiny* (1985), *O češtině v číslech* (1987), *Quantitative linguistics* (1992).

Vycházejí specializované slovníky vytvořené pomocí počítačů – **Eleonora Slavíčková:** *Retrográdní morfematický slovník češtiny* (1975) – a obsahující frekvenční charakteristiky slovní zásoby – **Marie Těšitelová**: *Frekvenční slovník současné české publicistiky* (1980); *Frekvenční slovník současné české administrativy* (1980); *Frekvenční slovník současné odborné češtiny* (1982); *Frekvenční slovník jazyka věcného stylu* (1983); J. Králík , M Těšitelová (1986): *Retrográdní slovník současné češtiny*).

Kromě toho byla věnována pozornost praxi i teorii v oblasti výzkumu univerzálních kvantitativních vlastností přirozeného jazyka (Luděk Hřebíček, Jan Králík, Marie Königová, Ludmila Uhlířová ad.; srov. např. Hřebíček (2002) aj. V roce 1994 byl založen časopis Mezinárodní asociace kvantitativní lingvistiky – *International Quantitative Linguistics Association (IQLA) Journal of Quantitative Linguistics*, kde zmínění odborníci také pravidelně uveřejňovali a uveřejňují výsledky své výzkumné práce.

Diferenciace akcentů v metodologických přístupech na straně jedné a nepříznivý politický vývoj na straně druhé vedl v letech 1968 – 1973 k tomu, že část badatelů (Ladislav Nebeský, Pavel Novák aj.) zůstala na FF UK, část (Petr Sgall, Eva Hajičová, Jarmila Panevová a další) založila Laboratoř algebraické lingvistiky FF UK, ale musela brzo přejít na MFF UK. Až v po roce 1989 byly založeny Ústav formální a aplikované lingvistiky (ÚFAL) na MFF UK a Ústav teoretické a komputační lingvistiky (ÚTKL) FF UK, tedy samostatná pracoviště

orientovaná výhradně na studium oboru matematické lingvistiky na UK (srov. dále). Všichni badatelé přesto po celou tuto dobu s určitými omezeními pokračovali ve výzkumné i pedagogické činnosti. Plodem jejich práce byla řada statí vycházejících v domácích i zahraničních odborných periodicích (*PBML*, *Slovo a slovesnost*, *Čs. informatika* aj.). Z monograficky zaměřených prací jmenujme alespoň **Petr Sgall – Ladislav Nebeský – Alla Goralčíková – Eva Hajičová**: *A functional approach to syntax* (1969), **Petr Sgall – Eva Hajičová – Eva Benešová**: *Topic, focus and generative semantics* (1973), **Petr Sgall – Eva Hajičová – Eva Buráňová**: *Aktuální členění věty v češtině* (1980), **Petr Sgall – Eva Hajičová – Jarmila Panevová**: *The meaning of the sentence in its semantic and pragmatic aspects* (1986), **Eva Hajičová**: *Negace a presupozice ve významové stavbě věty* (1975) a **Jarmila Panevové – Eva Benešová – Petr Sgall**: *Čas a modalita v češtině* (1971), **Jarmila Panevová**: *Formy a funkce ve stavbě české věty* (1980). (Podrobné hodnocení formálního přístupu v lingvistice a místa, které v něm zaujal ve druhé polovině 20. století český výzkum viz Hajičová – Panevová – Sgall, 1991.)

Přechod Karla Paly na katedru českého jazyka, slovanské a obecné jazykovědy filozofické fakulty Univerzity Jana Evangelisty Purkyně v r. 1964 byl podnětem pro experimenty v oboru matematické lingvistiky na univerzitě v Brně. Byly zaměřeny na syntaktickou a sémantickou analýzu přirozeného jazyka a rozvíjely se v širší spolupráci s odborníky jiných oborů, například s logikem Pavlem Maternou, který u nás uváděl práce českého logika Pavla Tichého (srov. Tichý, 1996; Svoboda – Jespersen – Cheyne, 2004), a s informatikem Jiřím Zlatuškou. Od druhé poloviny 70. let bylo možné díky technické spolupráci s Vysokým učením technickým v Brně a později s Ústavem výpočetní techniky brněnské univerzity ověřovat adekvátnost teoretických přístupů prvními experimenty s automatickou syntaktickou analýzou češtiny (programový systém Wander ve spolupráci s programátory Ústavu výpočetní techniky Miroslavem Benešovským, Martinem Šmídkem a Josefem Gerbrichem). Jistý průlom představují 80. léta. Na katedře českého jazyka obecné a srovnávací jazykovědy brněnské univerzity spolupracuje Karel Pala, Klára Osolsobě a Stanislav Franc na integrovaném morfologicko-syntaktickém analyzátoru *klara*, využívajícím jazyk Prolog a aparát DC gramatik (Definite Clause Grammars). Články publikované v řadě A *Sborníku prací filozofické fakulty brněnské univerzity* referují podrobněji o jednotlivých pokusech. Monograficky jsou výsledky spolupráce odborníků zabývajících se exaktními metodami ve vztahu k přirozenému jazyku zachyceny v monografii **Pavel Materna – Karel Pala – Jiří Zlatuška**: *Logická analýza přirozeného jazyka* (1989).

Jednou z aplikací NLP, spadající do oblasti language engineering (jazykového inženýrství), je jazyková podpora v textových editorech a sázecích systémech. Patří sem jazykové korektory pravopisných překlepů (tzv. spelling-checkery), korektory pro opravu gramatických, popřípadě stylistických chyb (grammar-checkery, style-checkery), tezaury nabízející uživateli řady významově blízkých slov, programy automatického dělení slov na konci řádku, popřípadě vícejazyčné slovníky, které autor textu může během práce s editorem otvírat jako samostatná „okna“ (nejde tedy o strojový překlad, ale o elektronickou podporu pro překladatele). Práce zaměřené na formální popis české morfologie (Hajič, 1994; Osolsobě, 1996) našly uplatnění v aplikacích zaměřených na automatickou korekci českých textů.

Do oblasti aplikací NLP patří také výzkum a vývoj automatického zpracování textu (automatická indexace, automatická tvorba terminologických tezaurů a automatický překlad). V této oblasti vznikla řada originálních systémů. Většina z nich byla vyvíjena jako projektové úkoly v rámci bývalé soustavy VTEI (Informační soustava vědeckých, technických a ekonomických informací) na specializovaných pracovištích nebo v jednotlivých oborových nebo odvětvových střediscích VTEI. V 70. letech 20. stol. se musíme zmínit alespoň o experimentech s automatickou indexací, které prováděl J. Janoš v OBIS při závodě Turbiny podniku Škoda Plzeň (Janoš, 1976) a za zmínku stojí také jednoduchá, ale účinná metoda automatické indexace AUTIS-AI vyvinutá na konci 80. let J. Hradilem v ODIS VTEI pro uhelný průmysl v Ostravě (Hradil, 1987; Rozkopal, 1994). Od začátku 70. let byl v rámci soustavy VTEI vyvíjen i systém SEMAN – SÉmantický ANalyzátor (srov. Smetáček, 1982, 1984, Uličný, 1987). Ten byl a dosud je používán k automatické popř. automatizované tvorbě tezaurových systémů. Na akademické půdě v rámci výzkumu katedry aplikované matematiky MFF UK byl vyvinut systém MOZAIKA (na Morfologickém Odvozování Založené Automatické Indexování Koherentními Agregáty) (srov. Kirschner, 1979, 1983). V souvislosti s rozpadem soustavy VTEI po roce 1989, resp. 1991, oba největší systémy (SEMAN a MOZAIKA), budované téměř 20 let, fakticky zanikly. Metodologie a technologie byly však prostřednictvím autorů těchto systémů alespoň částečně přeneseny do nových projektů. Ty jsou dnes často budovány na komerční bázi.V oblasti pojmového modelování a tvorby znalostních bází byl rozvinut velmi pozoruhodný projekt v Ústavu státu a práva AVČR, který je realizován v rámci právního informačního systému LEGSYS /LexGalaxy (srov. Kořenský – Cvrček – Novák, 1999, a též http://www.lwgsys.cz). Více informací o jednotlivých systémech i další bibliografické údaje, lze najít v přehledové studii (Schwarz, 2005, a http://www.ikaros.cz/Clanek.asp?ID=200303002)

Hovoříme-li o historii oborů zahrnovaných pod střešní název matematická lingvistika, pak nemůžeme opominout skutečnost, že COLING 1982 (9. mezinárodní konference komputační lingvistiky) se konala v Praze 5. 7. – 10. 7. 1982 na Karlově univerzitě. Předsedkyní organizačního výboru tohoto kongresu byla E. Hajičová a jeho konání bylo umožněno tím, že je díky Jánovi Horeckému oficiálně zaštítila Slovenská akademie věd. Nešlo tu pouze o uznání zdatnosti našich odborníků, kteří na domácí půdě přednášeli a hostili své zahraniční kolegy, ale i o to, že v době, kdy účast na mezinárodních fórech byla v naší části politicky rozděleného světa podstatně omezována, umožnilo konání konference prezentovat výsledky své práce lidem, kterým by k tomu jinak nebyla dána příležitost.

Připomeňme, že už koncem 19. století byly v reakci na dosavadní zkoumání jazyka z ryze historického hlediska položeny základy strukturalistické lingvistiky. Otcem tohoto hnutí byl Ženevan Ferdinand de Saussure.

Lingvistiku vracející se k Saussurovi lze charakterizovat s ohledem na její vědecko-teoretická východiska jako empirickou vědu, jejímž cílem je synchronní popis jazyka opírající se o analýzy empiricky uchopitelného materiálu jednotlivých jazyků. Z tohoto přístupu vzešla myšlenka textového korpusu. Korpusy dnes zahrnují jazykové jevy v podobě „masových dat", která lze uchovávat a zpracovávat pomocí počítačů. Tak dochází k úzkému propojení korpusově orientovaného výzkumu jazyka a počítačové lingvistiky. Z komputační lingvistiky se začíná postupně vydělovat korpusová lingvistika, lišící se od ní orientací na masové zpracování korpusových dat a na aplikace z něj plynoucí. Prvním moderním elektronickým korpusem byl *The Brown Corpus of Standard American English*, většinou uváděný pod názvem *Brown Corpus*. Vytvořili jej W. Nelson Francis a český rodák Henry Kučera. Ačkoliv z dnešního hlediska jde o korpus velmi malý (1 milion slovních tvarů), jednalo se o první korpus v dnešním slova smyslu (elektronický, složený ze vzorků vybraných z široké škály textů tak, aby byl dodržen požadavek reprezentativnosti korpusu). V roce 1967 shrnuli oba autoři výsledky analýz vycházejících z *Brown Copusu* v nyní již klasickém díle **Henry Kučera – Winthrop Nelson Francis**: *Computational Analysis of Present-Day American English.* Přestože byla korpusová lingvistika zpočátku vystavena značné kritice (zejména v 50. a 60. letech ze strany N. Chomského), stala se postupem času významným metodologickým proudem jazykovědy. Kromě mnoha korpusových projektů orientovaných na angličtinu se začínají budovat korpusy dalších jazyků. Velmi podrobné informace o korpusech jednotlivých jazyků a řadu dalších odkazů lze najít na http://www.athel.com/corpus.html. Od roku 1996 vychází *International Journal of Corpus Linguistics* (IJCL), přinášející širokou škálu názorů na roli korpusové lingvistiky ve výzkumu jazyka, počítačové lexikografii a NLP.

Od počátku 90. let 20. století se **korpusová lingvistika** dynamicky rozvíjí také u nás. Na jaře roku 1992 se sešla v Praze skupina badatelů (František Čermák, Jan Hajič, Eva Hajičová, Jan Králík, Karel Pala, Klára Osolsobě, Věra Schmiedtová ad.), kteří založili zájmové sdružení *Počítačový fond češtiny* (PFČ) (srov. podrobněji Čermák – Králík – Pala, 1992). Cílem tohoto sdružení bylo koordinovat úsilí a zajišťovat komunikaci a spolupráci odborníků, kteří mají zájem o počítačové zpracování českého jazyka. Posléze jejich snahy nabyly institucionalizované podoby. Prvním krokem byla grantová podpora (vůbec první grant nesl název „Počítačový korpus českých psaných textů", od roku 1993 zahrnoval spolupráci odborníků univerzity Karlovy v Praze, Masarykovy univerzity v Brně a Ústavu pro jazyk český, byl úspěšně ukončen v roce 1995). Klíčový význam pak mělo v r. 1994 založení samostatného pracoviště Ústavu Českého národního korpusu (http://ucnk.ff.cuni.cz) v čele s Františkem Čermákem (viz Korpusová lingvistika). Na další rozvoj bohemistiky má velký vliv budování rozsáhlých jazykových korpusů a korpusových nástrojů na straně jedné a „vytěžování" (mining) korpusů (využívání jazykových korpusů jako zdrojů informací o jazyce) na straně druhé. Nepochybný a netrpělivě očekávaný bude význam využívání korpusů pro počítačovou lexikografii; srov. Čermák (1999), Čermák – Klímová – Pala – Petkevič (2001). Prvním slovníkem založeným na korpusových datech v moderním slova smyslu je **František Čermák – Michal Křen**: *Frekvenční slovník češtiny* (2004).

V rámci korpusové lingvistiky se dnes termínem *korpus* označuje rozsáhlý (vymezení rozsahu korpusu je dáno účelem, k němuž se korpus buduje) soubor počítačově čitelných (MRF – Machine Readable Form) textů složený ze souvislých textových úseků vybraných podle jistých pravidel tak, aby reprezentovaly pokud možno jazyk jako celek v celé jeho pestrosti (reprezentativnost korpusu), který obsahuje standardní reference (anotace). Tyto reference zahrnují nejrůznější metatextové informace (vzhled textu, členění na kapitoly, odstavce, typografie, ale také údaje o typu textu – informace o autorovi, dataci, žánrovém zařazení atd.) a interpretace jednotek, z nichž je text složen (jazykové značky – tagy – anotace slovnědruhové, morfologické, syntaktické, sémantické, prozodické aj.). Iniciativu při budování standardního způsobu anotací převzala iniciativa TEI (Text Encoding Initiative). Jedná se o aktivitu sponzorovanou hlavními vědecky orientovanými asociacemi zabývajícími se využitím počítačů v humanitních vědách: ACL (Association for Computional Linguistics), ALLC (the Association for Literary and Linguistic Computing), ACH (the Association for Computers and Humanities). Cílem TEI je vytvoření standardní implementace pro operace s počítačově čitelnými texty. TEI za tímto účelem používá již existující formu značkovacího jazyka SGML (Standard Generalised Markup Language), popř. XML (EXtensible Markup Language). Značkovací jazyk je jakýkoli jazyk, který vkládá do textu značky vysvětlující význam nebo vzhled jednotlivých jeho částí. Vůbec první obecný značkovací jazyk byl Generalized Markup Language (GML). Na jeho základě byl vytvořen jazyk SGML, který se v minulosti stal jedním z nejrozšířenějších značkovacích jazyků. V dnešní době se nahrazuje jazykem XML – což je rozšiřitelný značkovací jazyk, jenž je zjednodušenou verzí jazyka SGML, vzniklou odstraněním chyb SGML a jeho modernizací. Vlastním příspěvkem TEI je detailní návod k použití příslušných standardů. Text jako celek se v TEI popisuje pomocí DTD (Document Type Description). Celá řada korpusových projektů přijala TEI za své. TEI vydává mnoho návodů pro kódování korpusových textů. EU založila dozorčí skupinu EAGLES (Expert Advisory Groups on Language Engineering Standards), která má za úkol sledovat různé evropské iniciativy a pomáhat jim. Jde o to vytvořit systém anotací, v němž by na jedné straně byla brána v úvahu specifika všech evropských jazyků a na straně druhé byla zachována jednota systému; srov. Čermák – Blatná (1995) a http://www.tei-c.org/.

V souvislosti s korpusovou lingvistikou se do středu zájmu dostávají aplikace NLP zaměřené na automatickou anotaci (značkování) velkých korpusů. Jsou vytvářeny počítačové programy pro automatické značkování – vkládání lingvistických informací do textu – velkých jazykových korpusů. Aplikace automatické morfologické analýzy se zaměřují na automatickou lemmatizaci (přiřazení základního tvaru tzv. lemmatu textovému tvaru slova), slovnědruhové značkování (POS – part of speech tagging) a přiřazování gramatických významů některých gramatických kategorií. Souhrnně se tento typ označuje jako gramatické značkování (tagování, popř. anotace, ale i lemmatizace). Vzhledem k tomu, že korpusy mají být zdrojem informací pro co nejširší okruh uživatelů, teoretikové korpusové lingvistiky se

zamýšlejí nad tím, jak by měla lingvistická informace vkládaná do textu vypadat a čemu by měla sloužit. (Některé zásady pro vytváření anotačních schémat formuloval Geoffrey Leech, 1993). Ruku v ruce s vývojem konkrétních anotačních nástrojů jdou tedy úvahy o mezích a možnostech, smyslu a účelu různých teoretických koncepcí vtělovaných do konkrétních interpretací jazykových jednotek, jimiž jsou tzv. značky (tagy) v případě slovnědruhových a gramatických značek, stromové struktury interpretující syntaktické vztahy s ohledem na sémantické, popřípadě pragmatické aspekty vyšších jednotek jazyka, značky postihující prozodické vlastnosti úseků textů, jeho sociolingvistické aj. charakteristiky atp. Klíčovou roli v tvorbě automatických korpusových analyzátorů hraje řešení problému disambiguace (zjednoznačnění homonymních jazykových jednotek na všech úrovních jazyka). Konkrétně o problémech spojených s morfologickým anotováním a následnou disambiguací českých korpusů srov. např. Hajič, 2004, Petkevič, 2001. K budování Pražského závislostního korpusu – Prague Dependency Treebank – PDT anotovaného na dvou syntaktických úrovních včetně aktuálního členění a hlavních typů koreference srov. např. Hajič, Hajičová, Panevová, Sgall, 1998, 2004, Hajičová, Panevová, Sgall, 2002.

Nesporný význam pro rozvoj oborů zaměřených na strojové zpracování přirozeného jazyka má od roku 1998 každoročně konaná konference Text, Speech and Dialogue (TDS), organizovaná Fakultou informatiky MU v Brně a Fakultou aplikovaných věd Západočeské univerzity v Plzni a od roku 2000 také International Speech Communication Association (ISCA). Je to jediná mezinárodní konference konaná pravidelně (každoročně) v České republice. Je setkáním odborníků z různých zemí a oblastí, jejichž společným zájmem je právě NLP (práce s jazykovými korpusy – texty, jejich přepis do strojově čitelné podoby (machine readable form – MRF), jazyková analýza, rozpoznávání, syntéza přirozeného jazyka v jeho psané i mluvené podobě, to vše na pozadí systémů zpracování přirozeného jazyka pomocí počítačů).

Velice dobrým zdrojem pro získání rychlé informace o světově uznávaných periodicích a konferencích je webová stránka digitálního archivu časopiseckých článků a příspěvků (papers) z mezinárodních konferencí oboru komputační lingvistiky zřízená Asociací komputační lingvistiky: (ACL – Association for Computational Linguistic – Anthology – A Digital Archive of Research Papers in Computational Linguistics http://acl.ldc.upenn.edu/). ACL je mezinárodní vědecké profesní sdružení lidí, kteří se zabývají problémy NLP. Ročně pořádá v létě konference ve významných centrech výzkumu zaměřeného na komputační lingvistiku. Asociace komputační lingvistiky má svůj časopis – *Computational Linguistics*. Asociace má americkou a evropskou sekci i řadu odborů pro jednotlivé problémové okruhy.

Na výše uvedené stránce lze najít *Computational Linguistics – CL Journal* (od r. 1980), *American Journal of Computational Linguistics* – ACL (od r. 1979), příspěvky z International Conference on Computational Linguistics – COLING (od r. 1965), Conference of the European Chapter of the Association for Computational Linguistics – EACL (od r. 1983), Conference on Applied Natural Language Processing – ANLP (od r. 1983), International Workshop on Natural Language Generation (od r. 1990). Řadu informací o jednotlivých světových pracovištích, asociacích, elektronických článcích i volně dostupném softwaru se zaměřením na NLP lze najít na http://www.ims.uni-stuttgart.de/info/FTPServer.html. Informace o dění v oblasti strojového překladu (bibliografie, konference, archív atd.) lze najít na stránkách The European Association for Machine Translation (EAMT) http://www.eamt.org/. Cenným zdrojem pro bibliografii v oblasti *computer science* je webová stránka http://www.informatik.uni-trier.de/~ley/db/. Účelně uspořádané informace o korpusech jednotlivých jazyků, archivech elektronických textů, korpusových nástrojích, el. dostupné literatuře i řadu bibliografických odkazů lze najít na http://www.athel.com/corpus.html.

Klára Osolsobě