**POČÍTAČOVÁ A KORPUSOVÁ LINGVISTIKA**

**1. Počítačová lingvistika (komputační – formální – kvantitativní lingvistika, NLP, language engineering). Vývoj ve světě a v ČR.**

* ***formální lingvistika*** – tj. odvětví lingvistiky zabývající se přesným zpracováním PJ (mluveného/psaného) pomocí matematických/informatických metod (matematická lingvistika), nebo pomocí pc programů (komputační lingvistika); usiluje o co nejpřesnější popis jazyk. struktur a dat ve všech jazykových plánech a mimoto se snaží jazyk modelovat pomocí logické formy; ***počátky***: spojeny s matematickým popisem PJ v 50. letech 20. st. (zvl. s rozvojem teorie formálních jazyků a automatů), zejm. pak boom s rozvojem jazykových korpusů
* ***komputační (pc) lingvistika*** – v širším smyslu vědní obor stojící na pomezí lingvistiky, matematiky a informatiky, v užším smyslu v popředí se jedná o zkoumání a zpracování PJ pomocí pc, pc technik a obecně informatiky
* ***kvantitativní lingvistika*** – 1 z oblastí matematické lingvistiky využívající matematické statistické metody (např. shluková analýza), teorie pravděpodobnosti (kvantitativní přístup – popisuje jistý jev pomocí proměnných, kt. lze vyjádřit čísly); počátky:
* ***jazykové inženýrství*** – jedná se o skloubení lingvistiky s informačními technologiemi, zabývá se počítačovými nástroji použitými při strojovém předkladu, extrakci informací, gramatických korektorech apod.

**2. Počítačové porozumění přirozenému jazyku a počítačové zpracování jednotlivých rovin přirozeného jazyka (morfologická, syntaktická, sémantická, slovotvorná).**

* **morfologická analýza** – ajka x poziční
* **syntaktická analýza –** synt x PDT
* **sémantická analýza** – vychází ze syntaktické analýzy a snaží se struktuře získané z parsingu dosadiz význam
* **slovotvorná analýza** – slovo – tj. řetězec znaků, derivace (připojení/odpojení) jisté části řetězce, Deriv, Derivanče

**3. Korpusová lingvistika – empirická věda. Historie korpusové lingvistiky. Definice korpusu v moderním slova smyslu. První korpusy. Chomského kritika korpusové lingvistiky.**

* ***definice korpusu*** – v obecné rovině se jedná o soubor dokladů autentického užití PJ (pro lingvst. analýzu a popis jazyka, jak mluveného/psaného); v historii lingvistiky se jednalo o označení pro jakýkoli lingvist. materiál, kt. vznikl ručním shromažďováním –> (50. léta 20. st.) vstup pc znamenal zásadní zlom, dky se elektronické korpusové projekty natolik rozšířily, že se pod pojmem ***korpus*** míní ***elektronicky uložený rozsáhlý soubor textů*** (pc zpracovaltený); obsahuje text, metaifnormace (autor, rok atd.), strukturu dokumentu, značkování
* ***korpusová lingvistika*** – zkoumá jazyk na základě korpusů (vznikají slovníky na základě korpusů apod.)
* ***vznik korpusové lingvistiky*** – v 19. st. s jiným významem; navázala na strukturalistické pojetí jazyka (synchronní přístup; myšlenka korpusu a požadavek na procedury, kt. mají odhalit pravidelnosti v jazyce); korpusová lingvistika s e-korpusy vznikla v Anglii, Brown corpus (Kučera a Francis)
* ***Chomského kritika k. l***.: vědy jsou obecně empirické (založené na pozorování přirozeně vyvíjejících se jevů) a racionalistické (přístupy zaměřeny na analýze uměle vyvolaných jevů); tvrdí, že korpus není pro lingvisty užitečný nástroj, neboť lingvista se má snažit modelovat jazykovou kompetence (langue; schopnost) a ne popisovat performanci (projevu, parole)

**4. Typy korpusů (kritéria – časové hledisko, druh zachycené komunikace, rozsah, žánrové hledisko, účel, jazyk). Specializované korpusy. Žákovské korpusy.**

* ***typy korpusů – kritéria*** (vznik korpusu by měl odpovídat účelu, pro kt. korpus vznikl): ***(1) z hlediska počtu obsažených jazyků***: jednojazyčný, vícejazyčný (paralelní; obsahují stejné texty ve více jazycích, zpravidla originál+jeho překlad); ***(2) z hlediska tématu***: obecný (budovány za účelem obecného popisu jazyka na všech jeho úrovních), specializovaný (vybudovány za speciálním cílem – např. studentské slovníky, nejfrekventovanější slova v určité oblasti apod.); ***(3) z hlediska modu***: mluvený, psaný, ***(4) z hlediska časového záběru***: synchronní (zabývají se kratším čas. obdobím; např. Brown korpus obsahující texty pouze z r. 1961), diachronní (zpracovává jazyk z hlediska delšího časového úseku –> lze pozorovat jisté vývojové změny); dále se vyčleňují tzv. všeobecné korpusy, kt. nevznikly za žádným blíže specifikovaným lingvistickým účelem (zpravidla slouží ke studiu sl. zásoby, gramatiky apod.); ***(5) text*** – zda se vkládají celé texty, či pouze jejich vzorky (kt. celý text vhodně reprezentují); (6) korpus otevřený (monitorovací)// uzavřený (většina)
* ***žákovské korpusy*** – patří mezi akviziční korpusy (tj. korpusy sloužící primárně ke studiu osvojování jazyka, sekundárně se mohou užít v didaktice jazyka – při tvorbě učebnic apod.; zkoumají rozvoj jazyka u dětí), zachycují užívání jazyka nerodilými mluvčími
* ***specializovaný korpus*** – vybudovány za speciálním cílem – např. studentské slovníky, nejfrekventovanější slova v určité oblasti apod.

**5. Budování korpusů. Získávání textů. Přepis, OCR metody, konverze dat. Značkování korpusů (strukturní značky, atributy). Metatextové anotace. Sociolingvistické značkování korpusů mluvených a korpusu soukromé korespondence.**

* ***budování korpusů*** – ***(1) získávání textů*** (většina textů v čnk získána v elektronické podobě přímo od nakladatelů na základě smluv; dále manuálním přepisem, skenováním, darem od autorů atd.), ***(2) zpracování textů*** (každý text se zaeviduje a vé své původní podobě je uložen do text. archivu, dále se všechny texty konvertují do jednotného formátu SGML, anotovat je a označkovat, takto jsou uloženy do text. banky a dále se s nimi pracuje jako se soubory; načež každý soubor je jednoznačně identifikován svým jménem a opatřen hlavičkou týkající se technických info konverzi; v této části se také z textů odstraňují části netext. povahy – obrázky, tabulky a odstraňují se duplicity; obecně ***konverze má 3 kroky***: z původní formátu do meziformátu, pak z tohoto meziformátu do dalšího meziformátu a z posle meziformátu do SGML a při posledním kroku je text morfologicky označkován) ***(3) evidence textů*** (do tabulek se zapisuje trojí info – bibliografie, formální charakteristika text. souboru – editor, v němž je text zapsán, administrativní info – jméno souboru, jméno pracovníka, kt. soubor převzal k anotaci apod.), ***(4) značkování*** (vnější/vnitřní anotace), (5) automatická morfologická analýz a disambiguace, (6) technické zabezpečení, (7) přístup ke korpusům
* konverze, nebo OCR technika, manuální opis textu do PC
* OCR metody – optické rozpoznávání znaků; pomocí scanneru
* značkování – připojení lingvistických, strukturní údajů a metatextových info (strukturní údaje: (1) strukturní jednotky: věta, odstavec atd., strukturní atributy– atributy strukturních jednotek –> název, autor dokumentu
* ***korpus soukromé korespondence*** – korpus soukromé korespondence; přepis ručně psaných dopis z let 1990–2004; všechny texty opatřeny zákl. sociolingvistickými informacemi o autorovi dopisu a adresátovi: pohlaví (žena – muž), věk (4 věkové skupiny), vzdělání (2 stupně)

**6. Lingvistické značkování. Tokenizace, lemmatizace, automatická morfologická analýza, disambiguace (statistické metody, pravidlová disambiguace, hybridní metody). Lingvistické problémy grafického zachycení korpusů mluveného jazyka (přepis a pravidla přepisu v mluvených korpusech budovaných v rámci ČNK). Stručná charakteristika olomouckého mluveného korpusu.**

* ***lemmatizace*** – lemmatizátor je program, kt. přiřadí každému sl. výskytu odpovídající lemma; taggery – gramatické kategorie (na základě pravděpodobností/pravidel); disambiguátor
* ***rozdíl mezi ORAL 2013 a ORAL2006 2008***: syntaktická interpeunkce nahrazena pauzovou, zančení překryvu 2 promluv mluvčích, pokrytí celého území (Čechy i Morava)
* ***OMK*** – vzniká od r. 2002, pod vedením p. POřízky FF UP v Olomouci, zaměřený na městskou olomouckou mluvu, podobně začal vznikat jako PMK a BMK, ale jednotl. sociolingvist. charakterisitky (věk, pohlaví) rozpracovány více a některá přepisovací pravidla byla změněna

**7. Korpusy v českém prostředí. ČNK – historie a současnost. Charakteristika jednotlivých korpusů pod ČNK (korpusy řady SYN, projekt DIAKORP a další). Charakteristika korpusů budovaných na FI MU (webové korpusy).**

**8. Korpusové manažery a další nástroje pro práci s jazykovými korpusy. Volba korpusu vhodného pro řešení lingvistického problému. Morfologické značkování českých korpusů (dva analyzátory – pražský a brněnský, poziční a atributový systém, shody a odlišnosti, porovnání tagsetů).**

* ***korpusový manažer*** – tj. aplikace pro práci s korpusovými daty (texty): vyhledávání konkrétních konkordancí, vyhledávání kolokací, vytváření zákl. frekvenčních statistik apod., dotaz se pokládá pomocí dotazovacího jazyka; pracuje buď lokálně (text. korpus je na konkrétním pc, zejm. v počátcích) nebo na principu klient-server (uživatelům ke přístupná pouze klientská část, kt. zajišťuje především uživatelské rozhraní (jak se zobrazují výsledky apod.), ale vlastní vyhledávání probíhá v serverové části, kde je uložen i korpus; Manatee (autor Pavel Rychlý) – klientská část: Sketch Engine (využívá word sketeches, kdy pro zákl. tvar slova sestaví jeho nejčastější vazby zachycující kolokační a gramatické chování tohoto slova + slova s podobnými slovními profily (auto – vůz; nemusí se jednat jen o synonyma, ale vyskytují se ve stejném kontextu) a rozdíly mezi srovnávanými sl. profily) , KonText, NoSketch Engine
* ***korpusové nástroje*** (slouží k různému způsobu vytěžování korpusových dat):

**9. Značkování syntaktické roviny jazyka. Pražský závislostní korpus. Projekty VALLEX a VERBALEX.**

* parsing – tj. syntaktické značkování
* PDT
* VALLEX (ÚFAL; Ústav formální a aplikované lingvistiky, UK) – valenční slovník českých sloves s komplexní sémanticko-syntaktickou analýzou, obsahuje valenční rámce se zákl. sémanticko-syntaktickou charakteristikou nejfrekventovanějších sloves, dodatečné gramatické informace informace týkající se modifikace (idiomatiky, reflexitvity); dostupný: v XML, PDF, HTML
* VerbaLex –

**10. Lexikální významy – slova a kolokace z hlediska automatických nástrojů, využití korpusů, elektronické slovníky (obsah, budování), lexikální databáze (WordNet, EuroWordNet, thesaury).**

* lexikální význam – – tj. historicky a společensky podmíněný jazyk. odraz mimojazykové skutečnosti přiřazený lexikální formě; zpravidla slova (jednotl. tvary lexému) mají týž lexikální význam, ale jiný gramatický
* lexikální databáze
* ***thesaurus*** (tezaurus) – tj. slovník, kt. zobrazuje ucelený obraz slovní zásoby; organizovaný pojmově, postupuje onomaziologicky (tj. od obsahu k formě); řadí LJ dle souvislostí do tříd, zobrazuje hierarchické (vertikální) a horizontální uposřádání těchto tříd –> např. synonyma x výkladové slovníky, třízene alfabeticky, sémaziolgocký přístup (neefektivní z hlediska času) vytváří se jejich elektronická verze
* ***WordNet*** – lexikální databáze, ve kt. jsou slova a sl. spojení seskupena do synonymických řad (synsetů) a jednotl. synsety jsou propojeny sémant. vztahy (nejčastěji používané sémant. vztahy: hypero-hoponymie, meronymie-holonymie –> vztah část-celek, žák-třída)
* EuroWordNet – snaha propojit všechny WordNety (aktuálně 70 jazyků, vč. č. WordNetu)

**1. Počítačová lingvistika (komputační – formální – kvantitativní lingvistika, NLP, language engineering). Vývoj ve světě a v ČR**

**Korpusová lingvistika** – nová větev lingvistiky, v níž se pracuje s korpusy uloženými v počítačích. Korpusy jsou v jazykovém inženýrství východiskem pro realistický základní výzkum ve formě relativně blízké přírodním vědám. Umožňují dělat věci, které dříve byly nepředstavitelné pro časovou náročnost a pracnost, systematičtější využití statistických a pravděpodobnostních přístupů.

Významné úkoly v NLP:

* analýza přirozeného jazyka – morfologická, syntaktická, sémantická
* generování přirozeného jazyka
* syntéza a rozpoznávání řeči
* strojový překlad (Machine translation)
* odpovídání na otázky (Question answering)
* získávání informací (Information retrieval)
* korektura textu (Spell-checking, Grammar checking)
* extrakce informací (Information extraction)
* výtah z textu (Text summarization)
* určení typu dokumentu (Text Classiﬁcation/Clustering)

**2. Počítačové porozumění přirozenému jazyku a počítačové zpracování jednotlivých rovin přirozeného jazyka (morfologická, syntaktická, sémantická, slovotvorná)**

**3. Korpusová lingvistika – empirická věda. Historie korpusové lingvistiky. Definice korpusu v moderním slova smyslu. První korpusy. Chomského kritika korpusové lingvistiky**

**4. Typy korpusů (kritéria – časové hledisko, druh zachycené komunikace, rozsah, žánrové hledisko, účel, jazyk). Specializované korpusy. Žákovské korpusy**

**5. Budování korpusů. Získávání textů. Přepis, OCR metody, konverze dat. Značkování korpusů (strukturní značky, atributy). Metatextové anotace. Sociolingvistické značkování korpusů mluvených a korpusu soukromé korespondence**

**6. Lingvistické značkování. Tokenizace, lemmatizace, automatická morfologická analýza, disambiguace (statistické metody, pravidlová disambiguace, hybridní metody). Lingvistické problémy grafického zachycení korpusů mluveného jazyka (přepis a pravidla přepisu v mluvených korpusech budovaných v rámci ČNK). Stručná charakteristika olomouckého mluveného korpusu**

**7. Korpusy v českém prostředí. ČNK – historie a současnost. Charakteristika jednotlivých korpusů pod ČNK (korpusy řady SYN, projekt DIAKORP a další). Charakteristika korpusů budovaných na FI MU (webové korpusy)**

**8. Korpusové manažery a další nástroje pro práci s jazykovými korpusy. Volba korpusu vhodného pro řešení lingvistického problému. Morfologické značkování českých korpusů (dva analyzátory – pražský a brněnský, poziční a atributový systém, shody a odlišnosti, porovnání tagsetů)**

**9. Značkování syntaktické roviny jazyka. Pražský závislostní korpus. Projekty VALLEX a VERBALEX**

**10. Lexikální významy – slova a kolokace z hlediska automatických nástrojů, využití korpusů, elektronické slovníky (obsah, budování), lexikální databáze (WordNet, EuroWordNet, thesaury)**