Využití korpusů pro lingvistická bádání

(volný překlad dle : ***Tony McEnery & Andrew Wilson:*** **Corpus Linguistics,** Edinburgh Teextbooks in Empirical Linguistics 1996, 1997)

Korpusy – zdroje empirických dat

V dnešní přednášce se budeme věnovat roli, kterou korpusy mohou sehrát a sehrávají ve zkoumání jazyka a při tzv. NLP, tedy snahách o tvorbu počítačového modelování jazyka. Podívejme se tedy nejprve na studium jazyka jako takového.

Jak jsme si již řekli, spočívá význam korpusů především v tom, že jsou zdrojem empirických (autorem neovlivněných, nezávislých) dat. Umožňují tak jazykovědci vyslovovat závěry, které jsou objektivně podloženy a neopírají se jen o subjektivní pozorování introspekci jedince. Použití empirických dat pro studium jazyka rovněž otevírá možnost studovat jazykové variety jako např. dialekty nebo starší stádia jazyka, které racionalistickým přístupem uchopitelné nejsou. Je patrné, že empirický výzkum je možný i bez korpusu. Celá řada lingvistů označují jako korpus data, která přísně vzato korpusem nejsou, protože neodpovídají všem požadavkům definice korpusu v úzkém terminologickém slova smyslu (neobsahují přesně definované vzorky, variety, atd.) Správně by měli říkat, že se opírají o sbírky textů (collections of texts). Můžeme tedy shrnout, že korpusová lingvistika nutně zahrnuje empirický přístup, ale empirický přístup nutně nepotřebuje korpus, tudíž ne každý, kdo přistupuje k jazyku empiricky je eo ipso korpusový lingvista.

V tom, co bude následovat, se podíváme na to, jakou roli mohou korpusy hrát v různých odvětvích lingvistického výzkumu. Zaměříme se na to, přoč jsou korpusová data důležitá právě v některých oblastech a jak může korpusově orientovaný výzkum přispět k rozvoji lingvistického bádání. Vy sami byste se měli zamyslet nad dalšími příklady, které vás napadnou (závěrečný návrh).

Korpusy a výzkum řeči

Zde existují dvě významné oblasti. Za prvé už samo budování korpusu mluvené řeči znamená shromáždění širokého výběru variant mluveného jazyka podle mluvčích lišících se věkem, původem, pohlavím, vzděláním. Vzorky pak jdou napříč různými žánry (konverzace na různá témata, beseda, přednáška, přednes, kázání atd.). Tato šíře záběru má dva klady. Šíře korpusu umožňuje snadnější generalizaci učiněných pozorování než omezený vzorek (reprezentativnost umožňuje generalizaci). Pokud můžeme vybírat z korpusu podle klíče, jímž je některý vzorek, můžeme zkoumat menší subkorpusy (jak mluví ženy, mládež, jak vypadají kázání) a studovat variety jazyka jednotlivě.

Druhou výhodou korpusu mluveného jazyka je, že zahrnuje vzorky přirozené řeči tak jak se mluví ve skutečnosti. Zde záleží především na tom, aby respondent, informátor mluvil přirozeně a nepřizpůsoboval, nemodifikoval svůj projev.

Vzhledem k tomu, že korpus mluveného jazyka (KMJ) zahrnuje většinou prosodické anotace,můžeme jej zpracovávat pomocí kvantitativních metod. Je možné postupovat dvěma způsoby. Za prvé je možné testovat na datech různé hypotézy, za druhé můžeme na základě pozorování dat hypotézu vytvořit, a pak ji na datech ověřit.

V obl. českých mluvených korpusů – málo až nic.

Korpusy a studium lexika

Lexikografové používaly empirická data dlouho před tím, než KL vznikla. Sběr dat pro slovníkové práce byl založen na sběru lístkových katalogů sestavovaných z příkladů lexikálních jednotek a jejich užití nalezených v různých zdrojích empirických dat (literárních aj. textech). Tak např. vznikl Oxford English Dictionary Samuela Johnsona v 19. stol. nebo PSJČ. Praxe opírající se o sběr citací stále v lexikografii pokračuje. Korpusy nicméně přispěly a přispívají

k rozvoji lexikografie velmi významným způsobem, neboť umožňují nejen lexikografům, ale všem lingvistům nový pohled na slovník.
Korpusy i libovolné sbírky textů umožňují vyhledání zadané lexikální jednotky v nejrůznějších případech jejího použití z velkých milionových korpusů během několika vteřin. Použití korpusů umožňuje rychlé rozšiřování a obohacování slovníku. Definice a výklad mohou být co nejbohatší až kompletní, velmi precizní. Příklady a doklady se díky počítačovému zpracování moho rychle řadit do smysluplných podskupin (kupř. pravý/levý kontext, abecední řazení atd.). Slovníky mohou obsahovat použití lexikální jednotky v různých jazykových varietách (regionální – srov. angličtina – britská, americká, australská,…), autor, žánr.
Při sestavování slovníků se nejčastěji používá tzv. otevřených monitorovacích korpusů, jak jsem se o nich zmínila ve druhé přednášce. V lexikografii hrají roli logicky proto, že se do nich s přílivem nových dat dostávají jednak nově vznikající lexikální jednotky, jednak nová užití existujících lexikálních jednotek. Tyto korpusy se vzhledem k neustálé proměně nehodí pro kvantitativní analýzy.
Ohromný význam mají korpusy (specializované) pro sestavování terminologických slovníků, kde právě souvýskyt a možnost jeho srovnání umožňují přesně definovat obsah termínu.
Tak jako význam slova je možné z korpusu vyčíst i údaje o jeho gramatických vlastnostech, zejména morfologii. Rychle a pružně můžeme zkoumat existenci, frekvenci a především rozvržení morfologických variant a produktivitu morfémů.

Korpusy a gramatika
Gramatika a syntax se často opírala o výzkum založený na datech korpusového typu. Korpusy jsou pro výzkum gramatiky důležité jednak proto, že přinášejí data, která je možné posuzovat z kvantitativního hlediska, jednak proto, že umožňují testování různých hypotéz a gramatických teorií.
V poslední čtvrtině dvacátého století se korpusově orientovaný výzkum gramatiky zaměřil především na kvantitativní analýzu, která pomohla jít za subjektivní tvrzení a odhalila skutečné rozdíly mezi obecným a zvláštním, frekventovaným a řídkým. V roce 1985 vydal Quirk a kol. gramatika Comprehensive Grammar of English Language založenou na korpusech.
Na korpusech lze frekvenční analýzou testovat celou řadu jevů popisovaných v klasických gramatikách a ověřit si minimálně jejich výskyt v korpusu, jejich frekvenci a při dalším zpracování význam (interpretaci) různých výskytů.
V minulých přednáškách jsme se několikrát zmínili o tom, že korpusový přístup stojí v opozici k tzv. racionalistickému přístupu. V oblasti zkoumání gramatiky se objevuje kombinace obou hledisek. Introspekcí získaná fakta se ověřují na materiálu korpusu a sleduje se tak obsah a rozsah jejich platnosti.

Korpusy a sémantika
Zmínili jsme se o tom, že v korpusu můžeme vyhledat jednotlivé výskyty slov, které nás zajímají a podívat se na ně, co v různých kontextech znamenají (lexikální sémantika – lexikografie). Korpusy mohou mít a mají svůj význam i pro obecnou sémantiku, protože umožňují výzkum významu založený na objektivním přístupu k jazyku.
Mindt (1991) ukázal, jak vytvořit objektivní kritéria pro odhalení významu lingvistických jednotek. Poukázal na to, že stanovení významu jazykové jednotky se většinou opírá o racionalistické přístupy (jak tomu rozumím). Na základě kvantitativních analýz různých významů lze ovšem ukázat objektivní význam různých jednotek opřený o empirickou evidenci. Sám zkoumal různé způsoby vyjadřování budoucnosti v angličtině.
Druhým příspěvkem KL v sémantice je možnost přesněji vymezit zhruba určené gramatické kategorie, přesně vymezit, která jednotka do kategorie patří a která nikoliv. Některé kategorie jsou uzavřené a mají pevně stanovené a stanovitelné hranice, jiné jsou otevřené a jejich hranice

se jen těžko vymezují. Korpusy mohou přispět tím, že místo toho, abychom se snažili postihnout hranici mezi kategoriemi v termínech patří do – nepatří do (inkluze – exkluze), postihneme rozdíl v termínech kvantitativních většinou – někdy – málokdy,, přičemž se můžeme opřít o výčet příkladů, které mohou sloužit jako vzory pro další neuvedené příklady, které se posléze vyskytnou.

Korpusy pragmatika a analýza diskurzu

V této oblasti je využití korpusů dosud poměrně málo zastoupeno. Pragmatika se definuje jako význam v kontextu. Vzhledem k tomu, že korpusy jsou sestavovány s menších vzorků, mohou být pro pragmatická bádání omezené.

Nicméně např. v PFG (Sgall, Hajičová – MFF UK) – zaměřuje se na zkoumání diskurzu a pracuje s korpusy.

Korpusy a sociolingvistika

Sociolingvistika sdílí s historickou jazykovědou, dialektologií a stylistikou to, že se odevždy opírá primárně o data, která můžeme v širším (ne úzce terminologickém) smyslu pokládat za korpusy (korpusová data). Vzhledem k tomu, že tato data nebyla sbírána primárně pro účely kvantitativního výzkumu, nebylo jejich sestavování podrobeno tak přísným požadavkům, které jsou kladeny na moderní korpus v úzkém slova smyslu (vzorky / reprezentativnost).

Většina sociolingvistických projektů a studií využívajících korpusy se zabývá zkoumáním lexika z hlediska tzv. genderové lingvistiky (vliv pohlaví autora textu na výběr lexikálních a jazykových jednotek).

Problém, na nějž se naráží a který by bylo možno odstranit, je nedostatečné zpřístupnění sociolingvistických informací v notacích korpusu a nedostatek sociolingvistického přístupu při sestavování vzorků. Sociolingvisticky anotované korpusy přinesou zajisté průlom v této abl. využití KL.

Korpusy a stylistika

Typické stylistické výzkumy se mnohem spíše orientují na výzkum jazyka jednotlivce (díla), než na zkoumání širokých variet jazyka. Stylistika se tedy spíše než o korpusy v pravém slova smyslu opírá o počítačově čitelné texty, které ovšem může analyzovat metodami používanými primárně KL.

Některé stylisty ovšem zajímá výzkum stylu určitého žánru, a pak přijdou ke slovu právě korpusy. (často žurnalistický styl).

Stylistika předpokládá, že autor vybírá z jazykových prostředků. Definice způsobu autorského výběru pak zakládá definici autorského stylu, tj. stupně v němž je výběr prostředků individuálním rysem. To lze zjišťovat kvantitativními metodami, které umožní objektivní srovnání.

Korpus je pak srovnávací bází.

Druhou oblastí, v níž se ve stylistice uplatní korpusy, je srovnání psaného a mluveného jazyka. Zde se nabízí dvě oblasti. Zajímavé je např. srovnat, jak se mluvený jazyk prezentuje v psaných textech – přímá řeč.

Korpusy představují výzvu pro zkoumání jazykové typologie (výzkum žánrů). Vzhledem k tomu, že korpusy zahrnují vzorky variet jazyka vybraných podle kritéria žánru, umožňují expost ověřit charakter typičnosti jazykových jevů pro příslušný žánr.

Korpusy a výuka jazyka a jazykovědy

Jazyková výuka odráží velmi dobře empirické a racionalistické přístupy jazykovědné teorie, o niž se opírají učební texty a metody. Mnohé učebnice jsou založeny na racionalistickém přístupu a obsahují texty, slovní zásobu, gramatické výklady, opřené o intuice autora. Jiné např.

projekt Collins- COBUILD se snaží opřít o korpusy a založit výklady a příklady na skutečném jazykovém materiálu.
Korpusy mohou navíc studentům studujícím cizí jazyky pomáhat přímo, jako zdroje, v nichž si mohou přímo hledat a ověřovat fakta, která vzhledem k nedokonalým znalostem jazyka, jenž se učí dobře neznají (lexikum, souvýskyt, obvyklost v rámci žánru atd.). Korpusy se pro mnohé vědce stávají přímým učebním materiálem. Pomáhají autorům učebnic při výběru slovní zásoby (frekvence), výkladu významu (kolokace), výkladu gramatiky atd.
Důležité je rovněž využití speciálních korpusů při sestavování učebních materiálů např. pro studenty nelingvisty (mediky, techniky), kteří potřebují zvládnout především jistou oblast jazyka (terminologii, odbornou frazeologii, pasivní zvládání jazyka – čtení odborných textů atd.).
Korpusy se používají nejen pro výuku jazyka , ale i pro výuku lingvistiky.
Jednou z oblastí využití korpusů je tzv. computer-assisted language learning. Na základě korpusů se vytvářejí různé softwarové nástroje sloužící k jazykové výuce.
Rozšíření vícejazyčných korpusů bude nadále sloužit pro výuku překladatelství a tlumočnictví (výzkum na univerzitě v Bologni – prof. Zanettin – článek v monografii Korpusová lingvistika).

Korpusy v diachronní lingvistice
Výzkum historických stádií jazyka je založen výhradně na korpusovém přístupu. Mrtvé jazyky a starší stádia vývoje živých jazyků jsou nám přístupny ve formě psaných památek, jejichž inventář je v podstatě omezen a vymezen. Jsou tedy z jistého hlediska reprezentativními korpusy. Je samozřejmě možné, že se někde objeví dosud neobjevený rukopis nebo neznámý nápis, nicméně celá řada problémů spojených se sestavováním synchronních korpusů se v diachronii neobjeví.
Otázka reprezentativního korpusu je poněkud jiná. V případě vzorků se používají podobné metody jako při výběru vzorků pro mluvený nebo dialektologický korpus. Roli hrají kritéria jako je doba vzniku a oblast původu vzorku.
I u diachronních korpusů platí, že rozsah zvyšuje význam korpusu, nicméně výsledky kvantitativních analýz se musí přijímat s větší opatrností, než u korpusů synchronních, protože máme k dispozici „jen to, co se dochovalo“, což nemusí být reprezentativní v tom smyslu, jako to, co si vybereme podle přísných kriterií pro korpus dnešního jazyka.
Helsinský korpus
DIAKORP

Korpusy v dialektologii
Dialektologické korpusy umožňují srovnání variet jazyka v jeho geografických varietách. Důležitá je především šíře vzorků po stránce obsahu a rozsahu a míra kompatibility vzorků z různých oblastí.
Velký význam má při výzkumu angličtiny existence korpusů britských, amerických, australské a novozélandské angličtiny, popř. korpusy anglických textů z oblastí bývalých kolonií.
V oblasti češtiny se uvažuje a projektuje převedení rozsáhlého materiálu dialektologických archívů do počítačově čitelné podoby.
O dialektologii lze říci, že jako jazykovědná disciplina pracuje s empirickými daty a že zpracování se zaměřuje na lexikum (primárně), gramatiku (morfologii), ale i např. syntax.

Korpusy a psycholingvistika
V centru zájmu psycholingvistiky stojí řešení otázek, kterak se produkuje jazyk v mysli. Zaměřuje se na měření takových hodnot, jako např. jak dlouho trvá nalezení hranic syntaktických jednotek při čtení, jak se pohybují oči a jak pracuje mozek.

Korpusy v psycholingvistice mohou přispět jako zdroje dat pro laboratorní experimenty. Frekvenční seznamy mohou pomoci k tomu, aby se psycholingvisté neblamovali např. zkoumají-li rychlost rozpoznání slov (u frekventovaných se vyšší rychlost dá předpokládat).
Svou roli mohou korpusy sehrát např. při zkoumání výskytu jazykových chyb v běžně mluveném jazyce (kdy je chyba věcí individuální a kdy obecnou, co z toho plyne pro její tolerovanost a tolerovatelnost atd.).
Také při výzkumu patologie jazyka hrají korpusy svou roli. Teprve poté, co jsou sestaveny korpusy jazyka tvořeného texty mluvčích s abnormalitami, můžeme postavit hypotézy o tom, jak u nich jazykové tvoření vlastně funguje.

Korpusy a kulturní studia
V lingvistice se dnes obecně má za to, že sociální podmínky ovlivňují charakter jazykových projevů. Vliv sociálních faktorů na volbu jazykových prostředků je oblast, kde se lingvistika stýká s oborem zvaným kulturní studia.
Zmiňuji studii srovnávající americkou a anglickou angličtinu, která na základě analýzy frekvence slovní zásoby došla k závěru, že Amerika je společnost maskulinní, militaristická, má větší vztah k mobilitě, kdežto britská společnost je femininní, mírná, stabilní.

Korpusy a sociální psychologie
Lingvisté nejsou jedinými uživateli korpusů. Korpus může představovat zdroj poznatků i pro odborníky z jiných oblastí. Pole výzkumu se díky korpusům se otvírá např. pro odborníky zkoumající sociální psychologii. Korpusy obsahují přirozená naturalistická data, která není možné získat v laboratoři.
Jedno z otázek, které patří do oblasti sociální psychologie je otázka jak a proč lidé vysvětlují některé věci. Projekt, který v roce 1987 Antaki a Naji provedli na London-Lund korpusu mluvené angličtiny se zaměřil právě na řešení tohoto problému a sice tak, že hledal výskyty textů následujících po příčinných a důsledkových spojkách.

Závěr
Na závěr bych chtěla shrnout čtyři výhody, které jsou společné korpusu a které z něj činí výhodnou bázi pro výzkum ve všech odvětvích výzkumu, o nichž jsme dnes hovořili.

1. Vzorkování a kvantitativní požadavky kladené na materiál korpusu z něj činí spolehlivou bázi pro generalizaci výsledků analýz.
2. Snadný přístup a počítačové zpracování urychlují získání dat všeho druhu.
3. Anotace poskytují data obohacená o informace, které by jinak bylo třeba pracně dodávat a dále s nimi pracovat.
4. Korpusy obsahují přirozená data (naturalistická data), neovlivněná laboratorními podmínkami, což zvyšuje objektivitu výsledků – nezávislost.