Historie korpusové lingvistiky

(volný překlad dle : ***Tony McEnery & Andrew Wilson:*** **Corpus Linguistics,** Edinburgh Teextbooks in Empirical Linguistics 1996, 1997 – repr. s. 1-19)

Za předchůdkyni korpusové lingvistiky lze pokládat každé lingvistické bádání založené na studiu vzorků – textů běžného, každodenního, neupraveného jazyka. Myšlenka korpusu je spjata se strukturalismem. Navzdory problémům, jimž korpusový přístup k jazyku musel čelit v 60. a 70. letech se v posledním desetiletí stala převažující metodologií studia jazyka.

KL obklopuje celá řada mýtů. Je to překvapivé s ohledem na to, že se jedná v prvé řadě o metodologický přístup. Mnohdy se KL vyčítá, že jejími výsledky se mohou stát truismy. Studiem korpusu lze samozřejmě odhalit „pravdy", k nimž lze dospět prostou úvahou – selským rozumem. Jiné námitky se týkají toho, že jazyk je nekonečný, zatímco korpus je vždy omezený a že tudíž nelze na základě korpusu vytvořit jeho úplný popis. Ruku na srdce. Co je to úplný popis. Lze tak složitý jev, jakým je přirozený jazyk, popsat úplně? To je opravdu filosofická otázka. Není bez zajímavosti si v této souvislosti připomenout, že ani v tzv. exaktních vědách, jimž se KL snaží přiblížit, nelze vyloučit experimentátora z experimentu. Je-li člověk, který programuje počítač pro analýzu korpusu omezený v nejširším smyslu slova, bude stejně tak omezená i výsledná analýza. To platí ale o libovolném popisu jazyka, nejen o tom, který je založen na analýze korpusu. Důležité ovšem je, abychom si všímali především aspektu metodologického (metoda – cesta – přístup).

Je korpusová lingvistika odvětvím lingvistiky?

Ano i ne. Není odvětvím tak jako např. morfologie, syntax, stylistika, je metodologickým přístupem k lingvistickému bádání. Opět opakujeme že je metodologií, jíž lze použít a jíž se rovněž s větším či menším úspěchem používá při zkoumání rozličných jazykových rovin, popřípadě v různých odvětvích lingvistiky. (Korpusová syntax, syntax založená na korpusu, sémantika,… sociolingvistika, dialektologie…)

Počátky korpusové lingvistiky

Za předchůdce KL můžeme považovat strukturalisty, kteří přišli s myšlenkou korpusu a procedury. Stručně řečeno je korpus text, jehož analýzou získáme jednotky, z nichž je vytvořen a odhalíme pravidla (procedury) podle nichž se kombinují. Připomeňme si alespoň analýzy literárních textů opírajících se pouze o samotné literární dílo, bez ohledu na jeho literární, historický aj. kontext (Mukařovský).

Několik slov o historii lingvistiky jako empirické vědy

Koncem 19. století byly v reakci na dosavadní zkoumání jazyka ryze z historického hlediska položeny základy strukturalistické lingvistiky. Ve středu jejího zájmu již není vysvětlování stavu jazyka z předešlého vývoje (diachronní hledisko), nýbrž popis stavu jazyka v určitém časovém bodě (synchronní hledisko). Pod pojmem jazyk (langue) se rozumí systém společný individuálním mluvčím, který je přítomen ve všech konkrétních promluvách (textech). Úkolem jazykovědy je především zkoumání a popis tohoto kolektivního systému.

Otcem tohoto hnutí byl, jak jistě víte, Ženevan Ferdinand de Saussure (1857-1913). Prominentní zástupci této školy byli i představitelé pražského lingvistického kroužku N. S. Trubeckoj, Roman Jakobson, V. Mathesius, J. Mukařovský... Dále pak kodaňská škola v čele s L. Hjelmslevem, v USA patřil k tomuto hnutí Edward Sapir, Leonard Bloomfield a C. F. Hocket a Zellig S. Harris. Rozkvět trval až do začátku 50. let.

Lingvistiku vracející se k Saussurovi lze charakterisovat s ohledem na její vědeckoteoretická východiska jakožto ***empirickou vědu.*** Jejím cílem je synchronní popis jazyka jakož to celku, který umožňuje individuu porozumět a vyjádřit se tak, aby mu bylo rozuměno. Jediným legitimním přístupem (metodou) je pozorování a analýza empiricky uchopitelného materiálu jednotlivých jazyků. Z tohoto přístupu vzešla ve strukturalistické lingvistice myšlenka **korpusu** a požadavek na **procedury**, které mají odhalit pravidelnosti ve zkoumaném jazykovém materiálu.

Myšlenka korpusu

Myšlenka **korpusu** vznikla z otázky, jak by měl být uchopen materiál příslušného zkoumaného jazyka. Je samozřejmé, že materiál jednotlivých jazyků je především všechno, co k určitému časovému okamžiku vyprodukují mluvčí příslušného jazyka, je to množina jazykových jevů vytvořená množinou mluvčích. Tento celek jazykových jevů se nazývá **korpus jazyka**.

V dnešní lingvistice má většinou jazyk mluvený primát ve vztahu k jazyku psanému. Lingvistika se věnuje v prvé řadě zkoumání mluveného jazyka a až na druhém místě bádání v oblasti jazyka psaného. Jazykové jevy se nejprve vyskytují v mluvených projevech, psaný text je odvozený pro potřeby zachycení mluveného textu prostřednictvím systému znaků.

V jazykovědné praxi v souvislosti s tím, co jsme řekli rozlišujeme:

*** korpus (korpusy) jevů mluveného jazyka**

*** korpus (korpusy) jevů psaného jazyka**

Pro oboje platí:

Korpus může být podle jazyka, který má být zkoumán, přehledný a uchopitelný, může být ale rovněž nepřehlédnutelně rozsáhlý, takže jeho popis je ryzí utopií. V takovém případě musí lingvista z téměř nekonečné množiny jazykových jevů vybrat část, o níž předpokládá, že s ohledem na otázky, jež si klade, je dostatečně reprezentativní. Hovoříme pak o **reprezentativním korpusu.**

Vytváření takovýchto korpusů je spojeno s mnoha problémy, které přináší stanovení kritérií pro reprezentativnost úseku vybraného z celku mluvených (psaných) textů.

V praxi se tyto problémy řeší na základě otázek, jež si klademe při zkoumání určitých jazykových jevů.

Tak můžeme např sledovat:

- jazyk určitého autora pro potřeby literární vědy

- historické stádium jazyka

-jazykové chování při komunikaci (rozhovoru) jisté sociální skupiny (např. mládeže - mládeže určitého věku, určitého geografického prostoru atd.)

V prvních dvou případech je problém reprezentativnosti vyřešen tím, že je celek materiálu k dispozici a lze jej většinou i dobře uchopit.

Horší je to ve třetím případě, kde je třeba uvažovat, kam stanovit mez pro reprezentativní korpus.

Pro problém reprezentativnosti korpusu je významné:

Typ Jazyk Forma Čas Počet textů Rozsah jaz.-aut. mluv.-ps.

Korpusy jazyka reprezentativní pro jazyk jsou korpusy vybrané podle jistých kriterií beroucích zřetel na komplex jazykových jevů příznačných pro zkoumaný jazyk.

Myšlenka korpusu byla ve strukturalismu spojena s další základní ideou, jíž bylo hledání vhodné procedury pro nalezení pravidel systému příslušného korpusu.

Procedury

Úkol odhalit a popsat systém pravidel daného jazykového korpusu, vedlo k otázce, které jednotky a vztahy konstituují hledaný systém. Podle principů strukturní lingvistiky jsou stanoveny procedury (přístupy), pomocí kterých jsou tyto jednotky a jejich vztahy určeny.

Ve svých nejpřísnějších formách žádá strukturní lingvistika, aby bylo možné vyvozovat bez jakýchkoliv předešlých znalostí závěry o jazykovém systému na základě pozorování jazykových jevů.

Z těchto úvah vychází myšlenka odhalovacích procedur (discovery procedures).

Označujeme tak přístupy jimiž mohou být ryze mechanicky odhaleny jazykové jevy a pravidla obsažená v korpusu.

Vůdčí ideou tohoto přístupu je představa, která mnohdy může odpovídat realitě, že člověk, který naslouchá textu, může odhalit pravidla, podle kterých je utvořen, na základě toho, jak je text rozdělen na jednotky a jaké mezi nimi platí vztahy. V souladu s tím spočívá částečná procedura ve vymezení jednotek textu a v určení jejich okolí. Jednotky jsou pak přiřazovány do tzv. **minimálních dvojic**, srovnává se jejich obsah. Touto metodou, o níž už nebudu víc hovořit, vzniknou konečně klasifikace, v nichž se odhalí pravidla řazení jazykových jednotek v textu.

Tyto typy odhalovacích procedur se nabízejí, aby byly automatizovány. Dosud však byly vyvinuty programy pouze pro oblast fonologie.

V lingvistickém bádání se daleko více vychází z toho, jazykovědec je s to, na základě své jazykové kompetence a prostřednictvím analýzy textů sloužících jako příklady zprostředkovat a modelovat jazykový systém. Pravidla, která během toho vytvoří mohou být přeměněna v procedury a použita na rozsáhlé korpusy textů. Tímto způsobem lze nejen popsat lingvistickou strukturu velkých textových korpusů, ale rovněž verifikovat vždy znovu lingvistické modely na jevech skutečného jazyka.

Myšlenka korpusu a myšlenka definování postavení jazykových jednotek v jazykovém systému a v jednotlivých textech prostřednictvím procedur, které se obě zrodily na poli strukturní jazykovědy, jsou základem pro některé přístupy ***computerové – počítačové lingvistiky***.

Korpusy totiž v sobě zahrnují jazykové jevy v podobě "masových dat". V dnešní době s nimi tudíž lze pomocí computerů lehce manipulovat. Procedury vyžadující

algoritmickou reprezentaci v jednotlivých krocích se dnes obvykle provádějí také pomocí počítačů.

Přesto lze jen zřídka použít k odhalení pravidel jazyka čisté odhalovací procedury. Častěji se počítá se znalostmi člověka, aby byly pomocí gramatik popsány jazykové zákonitosti obsažené v korpusu.

Lze ovšem jmenovat i celou řadu dalších předchůdců KL.

Jak se učíme muvit

Otázka jak se naučíme mluvit a s ní spojené studie dětské řeči jsou svého druhu korpusově orientované lingvistické práce, které se objevují již na konci XIX. stol. Analýzy rodičovských zápisků mapujících postupný rozvoj řečových schopností nejmenších dětí (Preyer 1889, Stern 1924) jsou dodnes používány (Ingram 1978) jako zdroj pro výzkum nabývání řečových schopností. Více se lze dozvědět v bibliografii, již T. McEnery uvádí v dopor. lit.

Studie zaměřené na spellování

Zkoumání rozložení hlásek v textu je velmi zajímavá otázka, jejíž řešení se prakticky odrazí např. v rozdílech klávesnic psacích strojů různých národů podle jazyka (francouzská klávesnice – kanadská). (Otázky spojené s zatížením hlásek/písmen řešili u nás např. strukturalisté Vachek – funkční zatížení fonémů, Trnka – otázky spojené s reformou těsnopisu).

Výuka jazyků a korpusy

Sestavování frekvenčních seznamů – výběr nejnutnější slovní zásoby. (srov. lit.)
(výběr SZ, gramatických jevů, frazeologie)
Korpusy mohou hrát při výuce jazyka důležitou roli. Dnes jsou využívány masově pro výuku angličtiny, ale i jiných jazyků. Informace lze získat na nejrůznějších www.
Korpus umožňuje přístup k jevům jazyka v té podobě, v níž se skutečně (aspoň někdy) přirozeně vyskytují. Jsou pomůckou pro učitele i žáky.

Role korpusů při výuce angličtiny
V uplynulých 15 letech se využívá korpusů pro výzkum toho, jak skutečně funguje jazyk. Výsledky se pak zpracovávají v jazykovém vyučování.
Slovníky založené na korpusech: Longman, Oxford, Collins
Gramatiky: Longman Grammar of Spoken and Written English, published in 2000
Největší pokrok: mluvený jazyk
DDL- data-driven learning (student je naveden k tomu, aby sám na zákl. matreriálu získaného z korpusu tvořil hypotézy o různých významech lexikálních, gramatických). Tím se u studentů rozvíjí deduktivní přístup, který když se jej naučí dobře ovládat jim pak pomůže v dalším studiu jazyka. (Nezavrhuji slovníčky a gramatická pravidla, ale pokládám za nemožné, aby se člověk naučil cizímu jazyku pouze z nich. V jisté fázi, když ukojíme svoji komunikační potřebu, většinou zakrňujeme ve studiu cizího jazyka. Dedukce nám v tomto krnění do určité míry, zapojíme-li vůli, může zabránit. Je dobrodružná, protože je tvůrčí.)
http://www.eli.ubc.ca/teachers/using_corpora.html

Srovnávací lingvistika
Studie zaměřené na porovnávání frekvencí slov v jedn. jazycích (40.l.)

Syntax a sémantika
Fries (1952) deskriptivní gramatika angličtiny založená na korpusu. Quirk (1985): Comprehensive Grammar of the English Language. Gougenheim (1956) – gramatický popis francouzštiny založený na frekvenčních analýzách korpusového materiálu.

V 50. letech zasadilo KL velkou ránu vystoupení N. Chomského.

S čím přišel Chomsky?
V dnešní době se slova korpus používá téměř „zaklinačsky". Věří se, že korpusový přístup je jedinou oprávněnou vědeckou metodologií v lingvistice. Je třeba ovšem vědět, že tomu tak nebylo vždy a že i lingvisté, kteří jsou dnes zastánci, či představiteli korpusového přístupu, neměli vždy stejný názor. Budeme se proto nyní zabývat chvíli historií.

Zároveň si také řekneme, proč námitky, které byly vznášeny, padly.
Tak tedy kupř. N. CH. (mluvíme o době 50. a 60. let, kdy začaly vznikat první počítačové korpusy) odvrátil sérií svých článků zájem od empirismu k racionalismu. Co se tím myslí? Je to něco specificky lingvistického? Na druhou otázku je třeba odpovědět ne. V každé vědě jsou přístupy založené na pozorování přirozeně se vyskytujících jevů (EMPIRICKÉ) a přístupy zaměřené na analýze uměle vyvolaných jevů (RACIONALISTICKÉ). Vyostříme-li to, můžeme říci, že racionalisté teoreticky vycházejí z umělých modelů chování a usuzují na základě introspekce. Takto rodilý mluvčí uvažuje o svém rodném jazyce a na základě těchto úvah vyvozuje teoretické závěry. Racionalistické teorie uvnitř lingvistiky se zakládají na teorii myšlení a jejich základním cílem je poznávací plausibilita. Jejich cílem není jen vytvoření modelu fungování jazyka ale tvrzení, že jazyk skutečně funguje podle jimi vytvořeného modelu. Zájem empiristů na druhé straně se upírá k pozorování dat vyskytujících se přirozeně a typicky prostřednictvím kopusu. (Věta x je větou jazyka, pokud se vyskytne v korpuusu).
Oba přístupy mají své klady a zápory. Na ně se ale podíváme později.
Chomsky posunul lingvistiku od abstraktního popisu jazyka k teorii, která bere zřetel na psychologické reality, na model jazyka jakožto prostředku poznání. Přitom ovšem zjevně devalvoval korpus jakožto zdroj lingvistických poznatků. Chomsky tvrdí, že korpus nemůže nikdy být pro jazykovědce užitečný nástroj proto, že lingvista se má snažit modelovat jazykovou competence (schopnost), a ne popisovat performance (projev). Pozdější Chomsky tyto své názory poněkud modifikoval, což lze vidět především z jeho prací z 80. let (1988). V této přednášce se ovšem budeme zabývat především jeho dřívějšími názory, které ovlivnily tříbení oboru KL.

Performace (parole) - projev competence (langue). Podle Ch. se mají lingvisté zajímat především o COMPETENCE. Performance je podle něj pouhým ubohým odleskem competence. Protože korpus je pouhou a navíc omezenou sbírkou performancí, je nutně ubohým zdrojem pro modelování jazykové competence.
Co se ale stane, když toto tvrzení nepřijmeme? Jak můžeme na základě zkoumání pouhé competence vyvodit, které jevy spadající do oblasti performance jsou

relevantní, které méně relevantní atp.Je nabíledni, že competence a její výzkum jsou nesmírně zajímavé, otázkou ovšem zůstává, co je důležité pro lingvistiku. (Je jistě nesmírně zajímavé vytvářet jazykové teorie, které objasňují, které jevy v jazykové kompetenci jsou „vrozené“. Jak to ale můžeme dokázat, nebudeme-li mít k dispozici reprezentativní množinu dětí, která se nesetkala s žádným projevem performance?).
Hlavním úkolem lingvisty je podle Chomského definice modelu jazykové kompetece (schopnosti). Je ovšem překvapivé, že odmítá korpus jakožto prostředek k dosažení tohoto cíle. Důsledně směřuje od empirismu k racionalismu, jazyk podle něj nemá být zkoumán pozorováním zvnějšku, nýbrž zevnitř.
Námitky vznášené proti průkopníkům KL byly do jisté míry oprávněné. Operovalo se totiž s představou kompletního korpusu, s představou, která vycházela z toho, že jazyk je konečná množina (korpus) vět, které lze spočítat, přehlédnout, popsat, atd. Spatřovat SOLE EXPLICANDUM jazykovědy v korpusu by bylo ovšem chybné. A právě některá Hockettova (1948) a Harrisova (1951) tvrzení vedla Chomského k jeho kritickým poznámkám na adresu KL.
Kdyby bylo možné popsat empiricky celek jazyka dostali bychom jistě objektivní popis. To je jistě pro každého vědce – lingvistu velmi přitažlivý cíl. Jazykověda by se tak postavila po bok fyzice a ostatním přírodním vědám. Je něco takového ovšem ve skutečnosti možné? Musíme i na základě Chomského kritiky. Je také třeba poznamenat, že Chomsky přesměroval lingvistická bádání od fonetiky k syntaxi a musíme přiznat i s ohledem na to, co jsme si tu už řekli, že reprezentativní korpus pro zkoumání fonetického systému je jistě snáze představitelný potažmo vytvořitelný než adekvátní korpus pro výzkum syntaxe. Je pravda, že zatímco počet fonémů je vymezitelný a omezený, je počet vět jazyka neomezený. Kdybyste si pozorně zapsali každou větu této přednášky, a pak si dali práci s tím, zda byste touž větu nalezli řekněme v potenciálním korpusu všech vět, které byly napsány a/nebo vysloveny před zahájením této přednášky, jistě byste tam alespoň většinu z nich nenašli, což svědčí o tom, že tyto věty jsou nové a že onen potenciální korpus dále rozšířily. Tento jev souvisí v posledku s možností aplikace tzv. rekursivních pravidel při tvorbě vyšších jazykových jednotek z jednotek nižších (opakování hlásek při tvoření slov, slov při tvoření vět, vět při tvoření souvětí atd.) Struktura věty obsahuje rekurzivní pravidla (slučovací souvětí, vedlejší věty vztažné atd. Korpus nemůže být tudíž nikdy sole explicandum přirozeného jazyka. Gramatická kompetence mluvčího je založena na znalosti syntaktických pravidel na naší syntaktické kompetenci, na jejímž základě jsme schopni vytvářet nekonečné množství vět. Data (korpus) obsahující pouze performance nikdy nemohou být vyčerpávající, úplná. Jazykové jevy nelze vyčíslit, žádný korpus nemůže tudíž odpovědně reprezentovat úhrn jazyka.
Chomsky (1959)
Každý přirozený korpus je nutně pokřivený (skewed). Některé věty se v něm nevyskytnou, protože jsou příliš primitivní, další proto, že jsou lživé, jiné zas proto, že jsou nezdvořilé. Má-li být korpus doopravdy přirozený, bude natolik pokřivený, že ani nebude nestát za zmínku.
Chomského pozorování je přesné. Korpusy jsou parciální ve dvojím smyslu slova. Jednak jsou neúplné. Obsahují pouze některé, a ne všechny věty jazyka. Druhým zdrojem jejich pokřivenosti je to, že jsou v nich primárně zastoupeny frekventované jevy a okrajové se do nich nedostanou, ačkoliv do jazyka bezpochyby patří stejnou měrou. Chomsky humorně tvrdí, že věta *„I live in New York . -Žiji v New Yorku.“* se v korpusu vyskytne jistě pravděpodobněji než věta *„I live in Dayton Ohio . -Žiji v Dayton Ohio.“* už jen proto, že v prvním ze jmenovaných měst žije bezpochyby více lidí, kteří mohou takové tvrzení pronést, než ve druhém. Co to ovšem říká o jazyce?

Nic. Proč má být tedy KL vůdčí metodologií, když jejím prostřednictvím docházíme k poznatkům, které nám nic o jazyce neříkají a introspekcí se dostaneme dále a rychleji? Korpus nás mnohdy vede k frustraci nad tím, kolik práce jsme vynaložili a jak malý je její výsledek. Fillmore (1992) se k tomuto tématu humorně vyjadřuje následujícími slovy:
Měl všechna potřebná primární fakta, která potřeboval, ve formě korpusu čítajícího zhruba zilion slovních tvarů a svůj cíl spatřoval v tom, že z těchto primárních fakt vyvodí fakta sekundární. Právě teď se zabývá tím, že bude určovat relativní frekvenci 11 slovních druhů (člen je zvl. SD) na první a druhé pozici ve větě.
To je vskutku nejen komická, ale i smutná představa. Můžeme z ní však vyvodit něco velice důležitého. Proč používat tak rozsáhlý korpus ke zjištění něčeho, co lze odpovědně dokázat i na korpusu nesrovnatelně menším (vezměme v úvahu slovosledná pravidla omezující slovní pořádek v angličtině).
Chomsky naopak spatřuje sole explicandum jazyka v rodilém mluvčím, který pozoruje svou jazykovou kompetenci. Nemůžeme samozřejmě použít hrubou sílu. Jak zjistíme, že nějaká věta je agramatická. To, že není v korpusu nemůže přece být validním argumentem pro takové tvrzení a naopak.
Shrňme si tedy argumenty proti používání korpusu:
Na základě korpusu modelujeme performanci, a ne kompetenci.
Úkolem lingvisty není podle Chomského výčet a popis performancí, nýbrž introspekce a vysvětlení jazykové competence.
I kdybychom viděli lingvistův úkol v enumeraci, není možná proto, že jazyk je nekonečný. Jakpak může být konečný korpus sole explicandum nekonečného jazyka?
Introspekci nelze tudíž vyloučit alespoň v konečném stádiu. Mohlo by se nám stát, že ztratíme ze zřetele jevy, které do jazyka také patří.
Tyto argumenty daly KL hodně zabrat.
Nicméně to, co zde bylo o N. Chomském řečeno, nelze brát jako jediný zdroj informací o jeho lingvistických přístupech.

Co se namítalo dále?
První práce s velkými korpusy byly vzhledem k tomu, že se prováděly ručně velmi časově náročné (vyžadovaly mnoho člověko-hodin, jak se dnes říká) a navíc vzhledem lidskému faktoru obsahovaly množství chyb. Práce s nimi byla drahá a málo efektivní. Před vznikem počítačově uložených a zpracovávaných korpusů se celá řada námitek proti KL jeví naprosto oprávněna.
Počítače změnily zásadně KL, která se díky nim stala nejdůležitější oblastí rozvoje počítačové lingvistiky.

Jaký smysl má tato přednáška?
V naší přednášce se ovšem nebudeme zabývat minulostí, ale současným stavem.
Kritika ovšem neustává ani dnes. Všeobecně se má za to, že korpusový přístup je nosný v oblasti fonetiky, v oblasti výzkumu získávání jazykových schopností dětí, ve zpracování frekventovaně se vyskytujících jevů. (To nijak nepřekvapí – u dětí je introspekce nepravděpodobná až nemožná. Introspekci můžeme provozovat, až když se vyvine metalingvistický aparát).
Korpusová lexikografie.

V 60. a 70. letech se sice KL rozvíjela, byla však pouhou okrajovou metodologií. Spolu s Chomského kritikou odhalující nevýhody datově orientovaného přístupu ke zkoumání jazyka se ovšem zároveň potvrdily i jeho přednosti.
K objektivním datům má přístup každý (veřejná kontrola privátních tvrzení). Veřejná verifikace tvrzení založených na korpusu je evidentní.
Umělá data jsou umělá. Svědčí o tom důkazy, které podal např. Sampson (1992), když ukázal, že introspektivisté se velice často zabývají jevy, které v korpusu nejsou typické, z čehož můžeme soudit, že jsou nejen okrajové, ale dokonce výjimečné. Tím se vyvrací do jisté míry Chomského výtka na adresu korpusu, že jsou věty, které tam nenajdeme. To je jistě pravda. Otázkou ovšem zůstává, jaká je jejich důležitost? Jistě neplatí, že nefrekventované rovná se nedůležité, nicméně nelze tento aspekt ztrácet ze zřetele.
Údaje o frekvence jsou jedním z nejsilnějších argumentů pro korpus, protože jedině on je zdrojem pro jejich získání. Všimněme si, jak často se necháme svést k formulacím s použitím kvantifikujících výrazů typu „většinou“, „především“, „zejména“. Ověříme-li skutečnou platnost některých z těchto tvrzení na korpusovém materiálu, lehce je vyvrátím prostou frekvenční analýzou.
Předností korpusu nejsou ovšem jenom frekvenční analýzy. Inspirativní jsou především obhajoby korpusu z pera G. Leeche, který dokazuje, že KL je nejsilnější metodologií z vědeckého hlediska právě pro objektivní ověřitelnost (verifikaci) výsledků. To nelze přiznat metodám založeným na introspekci.

Vraťme se ale ještě k jedné z námitek Chomského, která tvrdí, že většina performancí (korpusů) jsou agramatické. Zdá se ale, že to není úplně pravdivé tvrzení.

Korpus je tedy jedinečným zdrojem kvantitativních dat. Chomský poznamenává, že ta ale nejsou ve středu lingvistova zájmu. Opak dokázal např. Svartvik (1966). Kvantitativní data mohou odhalit zajímavá fakta a přinést mnoho užitečného jak pro teorii, tak i pro praxi.
Druhý okruh námitek vznášených proti KL odstranilo, jak jsme si již naznačili, spojení KL a počítačové lingvistiky.
Dnes pod pojmem korpus rozumíme velké sbírky materiálu přirozeného jazyka, t.j. texty, které vznikly bez vnějších zásahů lingvistů v psané nebo mluvené podobě a jsou uchovávány v elektronické formě. Umožňují elektronicky a tedy systematicky zkoumat jazyk v jeho přirozeném tvaru a nabývají tak čím dál tím více na významu především pro vědce lingvisty, nelingvisty ale i pro všechny zájemce, kteří chtějí proniknout k tajům jazyka. Termín korpus je tudíž identický s představou strojově čitelného korpusu. Počítače umožňují nejen bezpečné uložení korpusů, ale i jejich rychlé a bezchybné mechanické zpracovávání, prohledávání, třídění a počítání různých jevů. Úlohu počítačů v korpusové lingviste nelze ani přeceňovat, ani podceňovat. Záměrně jsem proto volila slovo „mechanické“ zpracování. Všimneme-li si ale právě toho, jak je slabý - pomalý, k chybám náchylný člověk vykonávající mechanickou práci s korpusem, je počítač naprosto bezkonkurenční pomoc. Musíme si ale rovněž připomenout, že od mechanických nebo z mechanických operací lze jít dále a i zde může hrát počítač svou roli. Celá řada pokusů jsou sice „slepé uličky“, ale i ty patří k vědeckému bádání.
Počítač prostřednictvím speciálních programů, tzv. korpusových manažerů, umožňuje rychlé a kompletní hledání a vyhledání definovatelných jazykových jednotek a jejich zobrazení, uložení a další zpracování, např. různé typy třídění a numerických operací. Konkordance, které jsou dnes běžně uznávanou formou

výsledku vyhledávání umožňují podívat se na hledaný jev z širší kontextově zapojené perspektivy.

Na závěr bychom si při zvážení námitek vznášených proti korpusům a důkazů pro ně měli uvědomit platnost obojího. Je patrné, že oba přístupy mají v lingvistice své místo. Nejdůležitější, co si máme uvědomit je to, že se vzájemně nevylučují a že možná nejlepší výsledky by mohla přinést kombinace obou.

Vývoj KL v době její nepopularity v 60. a 70. letech

Mohlo by se zdát, že KL má své počátky v strukturalismu, ty pak sahají do 50. let, kdy byla smetena se stolu Chomskym a až v 80. letech s prudkým vývojem počítačové technologie zaznamenala své znovuoživení. Takový pohled by byl ovšem poněkud zkreslující. I v 60. a 70. letech pokračovali někteří lingvisté přes námitky, výtky a posměšky dále v práci s korpusy a připravovali tak půdu pro dnešní prudký rozvoj tohoto odvětví. Quirk (1960) naplánoval a vytvořil korpus SEU (Surrway of English Usage). V témže roce začali Francis a Kucera prracovat na dosud známem Brown-Corrpus, jehož dokončení jim trvalo téměř dvacet let. V roce 1975 začal Jan Svartvik práci na London-Lund Corpus. Počítače začaly postupně nabývat na významu Svartvik vytvořil elektronickou podobu SEU, který je, jak tvrdí Leech, dodnes nepřekonaným zdrojem studia mluvené Angličtiny. Brown-Corpus je dostupný v počítačové podobě.

Proč vlastně lingvisté přese všechnu kritiku myšlenku na korpus nezavrhli? Je nanejvýš pravděpodobné, že si mnozí uvědomili, že technický rozvoj počítačů pomůže celou řadu námitek proti KL eliminovat.

Co bychom si měli zapamatovat?

Korpus a lingvistická intuice nestojí proti sobě, nýbrž mají se vzájemně doplňovat.

Ne antagonismus, ale komplementarita obou přístupů. Dovolím si tudíž na závěr své přednášky ocitovat Fillmora (1992):

Nemyslím si, že by kterýkoliv korpus, ať by byl sebevětší, mohl obsahovat informace o všech oblastech anglického slovníku a gramatiky, které bych rád zkoumal…[ale] každý korpus, který jsem měl příležitost zkoumat, ať už byl sebemenší, mne překvapil tím, že jsem v něm našel věci, na něž bych jinak nebyl přišel. Můj závěr je ten, že lingvistika potřebuje oba přístupy.