Kvantitativní data

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

Úvod

Kvantitativních dat, jejichž zdrojem se může stát korpus bez velkých problémů, jsou empirickými daty, k nimž, jak jsme si řekli v první přednášce našeho běhu, je třeba přistupovat s maximální opatrností. Připomeňme si jenom námitky vznášené proti korpusové lingvistice v 50. a 60. l. N. Chomskym. Vzpomeňme na poukazy a na výsledky, k nimž lze pomocí kvantitativních metod v korpusové lingvistice dojít a které jsou buď truismy, a nebo k nim můžeme stejně dobře dojít aniž bychom zkoumali korpus, zkrátka a dobře za použití selského rozumu (common-sence)

V minulé přednášce jsme hovořili o tom, jakých zásad je třeba se přidržet, abychom získali maximálně reprezentativní korpus, jak se pracuje se vzorky atp. Korpus v dnešním smyslu by měl být nesrovnatelně lepším zdrojem pro studium kvantitativních údajů, než veškeré jiné zdroje. Nicméně korpus nemusí být pouze zdrojem pro kvantitativní analýzy.

Kvalitativní a kvantitativní analýza.

Frekvenční seznam – slovních tvarů – slov – příp. dalších jednotek PJ.

Rozdíl mezi kvantitativní a kvalitativní analýzou korpusu spočívá v tom, že kvantitativní data přirozeně čerpatelná z korpusových textů nejsou součástí lingvistických rysů, které se datům přiřazují. Jsou pouze bází pro analýzu, která musí pokračovat dále. V kvalitativně zaměřené analýze jsou málo frekventované jevy zkoumány se stejnou pozorností jako jevy silně frekventované. Cílem analýzy korpusu není konstatování obvyklých a řídkých jevů v jazyce, nýbrž detailní popis jazyka jako celku.

Uveďme si příklad, kdy kvantitativní analýza neodhalí kvalitativní rozdíly. V korpusu nalezneme adj. rudý s určitou frekvencí. Prostá frekvenční analýza nám ovšem neodhalí, zda je toto adjektivum užito ve významu barevné kvality, nebo politické kvality. Existují cíle, k jejichž naplnění lze dojít, aniž bychom podobné rozlišení („jemnost“) potřebovali. K jejich splnění stačí kvantitativní analýza. Mnohdy se ale s podobnými výsledky spokojit nemůžeme, a pak se neobejdeme bez analýzy kvalitativní.

Kvalitativní analýza přináší tedy bohatší a přesnější výsledky, zatímco kvantitativní analýza podává statisticky spolehlivé zevšeobecnitelné výsledky, které mohou ovšem být do jisté míry zkreslující. V přírodních vědách se objevují metody kombinující oba přístupy a korpusová lingvistika se snaží o podobné řešení.

Reprezentativnost korpusu

Jak jsme se už zmínili v první přednášce, když jsme hovořili o kritice N. Chomského, který předhazoval KL její omezenost danou především omezeností korpusu oproti jazyku v jeho celistvosti, je třeba klást maximální důraz na reprezentativnost korpusu. Každý korpus je pouze vzorkem jazyka není jazykem v jeho úplnosti. Proto se může snadno stát, že v korpusu převládnou jevy okrajové a jsou potlačeny jevy centrální a že tudíž nebude spolehlivým ani jakožto zdroj kvantitativních údajů. Musíme si ovšem uvědomit, že v době (zač. 50.l.), kdy N. CH. svoji kritiku vznesl, existovaly pouze velmi malé korpusy. To samozřejmě souviselo s tím, že korpusy byly vytvářeny a zpracovávány především ručně. Dnes, kdy máme k dispozici výkonné počítače, ne kterých je možné ukládat, uchovávat a zpracovávat milionové korpusy, přestávají dřívější námitky platit.

V diskusích, jak docílit maximální reprezentativnosti korpusu, bylo zdůrazněno, že je třeba mít na zřeteli to, že vzorky představují širší populaci. Tuto populaci je nejdříve potřeba přesně vymezit, a pak z ní vybírat vzorky. K tomu, jak příslušnou populaci vymezit mohou existovat

různé přístupy. Chceme např. zkoumat veškeré texty vzniklé v určitém roce (vymezení populace). Ty mohou být získány a) z bibliografického indexu (LOB) b) ze všech akvizic velké knihovny (BROWN).
Jinak je samozřejmě třeba postupovat při sběru vzorků neformálních jazykových projevů (mluvený jazyk, soukromá korespondence), u nichž neexistují žádné zdroje mapující nebo zaznamenávající tyto texty, které ovšem do korpusu nepochybně patří. V těchto případech užívá KL tzv. demografické vzorky (podobně jako jiné spol. vědy – psychologie, sociologie), kdy předem definuje okruh tzv. informátorů (termín z dialektologického výzkumu). Definuje výběr informátora na základě jeho věku, pohlaví, oblasti, z níž pochází, dosaženého vzdělání atd. Tímto způsobem byly např. vybrány vzorky mluvené angličtiny PMK, MLUV).
Druhým aspektem vzorkování je potřeba stratifikace vzorků. Vzorky by měli v určité proporci pokrývat to, co bychom mohli široce nazvat žánry (žurnalistické texty, beletrie – růz. žánry, odborné a popularizační texty,..). Na tomto místě musíme ovšem upozornit na to, že už sama klasifikace žánrů a zařazení textů do jakékoliv klasifikační soustavy je věcí autorské interpretace, která je nutně široce vzato subjektivní, podmíněná teoretickými východisky, a tudíž ne absolutní a absolutizovatelná.
Dalším požadavkem při vzorkování je určení délky vzorku a počtu vzorků. Studie ukázaly, že zatímco frekventované jezy jazyka se jsou distribuovány v rámci textu stabilně, z čehož plyne, že k jejich zachycení stačí kratší vzorky, řídké jevy se objevují víceméně náhodně rozhodně nestabilně, což je důvodem toho, že nemusí být zachytitelné ani rozsáhlými vzorky. Tento druhý aspekt opodstatňuje odpovídající aspekty kritických postojů vůči korpusům.
Stanovení optimální délky vzorku a počtu vzorků je tudíž věcí nanejvýš problematickou. Bylo by totiž potřeba u každého vzorku vypočítat parametry, které nelze vypočítat pro korpus jako ideální celek. Těmito parametry je standardní odchylka každého jednotlivého rysu a tolerovatelná chyba, která se bude měnit v souladu s obecnou frekvencí rysů. Problematické je to především proto, že korpus se buduje proto, aby se na něm zkoumala celá řada jednotlivých rysů a nejen rys jeden. Biber proto navrhuje nepříliš konzervativní řešení. Výpočet by se měl řídit vždy rysem, který vykazuje maximální variabilitu. Tím by se měly pokrýt i ostatní, méně varírující rysy.
Pokud budeme při budování korpusu používat přísně tytéž statistické metody, dosáhneme přinejmenším toho, že budeme mít korpusy stejné míry reprezentativnosti.

Zpracování dat kvantitativními metodami

Ukázali jsme si, jak v KL jde kvantitativní analýza ruku v ruce s analýzou kvalitativní. Nyní se budeme zabývat alespoň v přehledu některými běžně užívanými technikami matematické statistiky, které v rámci KL následují za prostým počítáním frekvenčních výskytů jazykových jevů obsažených v korpusu. Díky těmto metodám se lingvisté snaží získat z korpusů nejen prostá kvantitativní data, ale dojít k interpretaci jejich závažnosti, a to pomocí exaktních matematicky ověřených postupů.
Jsou to např. metody, při jejichž užití je možné brát zřetel na takové okolnosti, jako je typ okolí jednotky (kolokace), vzorku (žánr) atd. Uvedeme krátce pouze ty metody, které jsou pro práci v KL nejobecněji užívané. Nejdříve ale musíme uvést dvě připomínky. Za prvé náš přehled je pouze omezený a snaží se bez zacházení do matematických podrobností pouze naznačit, jak některé používané metody fungují (nejsem matematický statistik a úvod do mat. stat. není cílem naší přednášky). Za druhé bych chtěla říci, že pokud by někdo hledal nějaké další poučení, mohu doporučit literaturu (Statistics for Corpus Linguistics v řadě edinburských učebnic empirické lingvistiky, internet).

Frekvenční analýza

Nejrozšířenější metodou při kvantitativním zpracování jazykového materiálu je frekvenční analýza, tedy vlastně matematické sečítání počtu jednotek (tokens) nebo v případě klasifikovaných jednotek typů. Tak můžeme získá frekvenční seznam slovních tvarů, slovních druhů, gramatických významů atd. U neanotovaného textu se většinou při použití běžně dostupných metod můžeme dopracovat k hrubším výsledkům než u textu anotovaného. Představme si, že budeme hledat slovo student. Pokud dotaz zadáme tak, že se ptáme na slovo student, získáme pouze řetězec písmen s-t-u-d-e-n-t. V případě, že chceme všechny textové podoby (studentovi, studentem, studentech, studentu, studenti, studentů…), budeme muset buď dotaz formulovat poněkud sofistikovaněji, nebo mít k dispozici označkovaný korpus, kde bude provedena lemmatizace. Grramatické značkování nám pak pomůže k tomu formulovat dotaz ještě obecněji, a sice např. zeptat se na všechny výskyty slova student v dat. sg.

Proporcionalita

Prosté počítání frekvencí má při korpusově orientovaném výzkumu důležité místo a často se používá, byť jen jako první krok další analýzy. Hlavní nevýhodou prostých frekvenčních výpočtů je, že výsledky, které jimi získáme, se mohou značně lišit v případě, kdy jeden a týž jev spočítáme v různých korpusech (např. psaném a mluveném). Stojíme pak před otázkou, jak získané výsledky porovnat a s porovnání vyvodit závěry. První problém nastane, když máme výsledky ze dvou korpusů, které nejsou stejně velké (jeden má 50 000 tokens, druhý 1 mil.) Dostáváme se k tomu, že potřebujeme nějakým způsobem započítat proporcionalitu do výsledků. Musíme jít za prostý výpočet aritmetické frekvence a vypočítat frekvenci jako procento z celkového počtu tokens v korpusu. Až teprve výsledek srovnání percentuálního zastoupení nám může říci něco spolehlivého.

Např. zjistíme, že jev A se v korpusu psaného jazyka o 1 mil. slovních tvarů vyskytuje 500 krát a v korpusu mluveného jazyka o 100 000 slovních tvarů 50 krát. Vypočítáme percentuální výskyt, a to takto:

mluvený korpus (50: 100 000) x 100 =0,05%
psaný korpus (500: 1000 000) x 100=0,05%

V obou případech nám vyjde stejný výsledek. Vypočítali jsme, že s ohledem na různost proporcí vzorků je frekvence stejná.

Lze určovat různé typy proporcí, nicméně vždy se vychází z poměru mezi velikostí vzorku a počtem výskytů.

ratio=počet výskytů typů /počet výskytů tokens v celém vzorku

Testování významnosti výsledků frekvenčních analýz

Podívejme se nyní na některé metody, které můžeme použít pro ověření významnosti výsledků frekvenční analýzy. V učebnici T. McEneryho a A. Wilsona je uveden tento příklad. Bude nás třeba zajímat, srovnání výskytu slovesa říci v latinském překladu dvou evangelií (Mat. a Jan.). Prostou frekvenční analýzou zjistíme, že u Mat. se objeví tvar dicit (říká) 46 krát a tvar dixit (řekl) 107 krát, zatímco u Jana najdeme tvar dicit 118 krát a tvar dixit 119 krát. Sestavíme si následující tabulku:

	dicit	dixit
Mat.	46	107
Jan.	118	119

Na první pohled by se nám mohlo zdát, že Jan používá častěji sloveso říci, než Mat. Jak ale zjistíme, jaký statistický význam má toto naše pozorování? Jak zjistíme, zda výsledek není jen náhodný?
Obvyklými testy používanými v KL jsou chi-squared test, t-test a Wilcoxons rank-sum test.
My si ukážeme chi-squared test, protože se nejčastěji v KL užívá vzhledem k tomu, že je a) vykazuje větší míru citlivosti při aplikaci na lingvist. data, b) nezahrnuje předpoklad, že data jsou normálně distribuována, což je předpoklad, který, jak víme, v přirozeném jazyce neplatí, c) v tabulce malých rozměrů 2x2 je výpočet tak snadný, že se obejdeme bez statistického software a můžeme si výsledek spočítat sami. Hlavní nevýhodou je ovšem malá spolehlivost u malých frekvencí.
Chi-squared test je založen na srovnání frekvence nalezené v textu a předpokládané (očekávané) frekvence. Čím jsou si obě čísla bližší, tím je pravděpodobnější, že frekvence je věcí náhody. čím je naopak rozdíl větší, tím je pravděpodobnější, že frekvence je signifikantní a je ovlivněna jinými faktory, než náhodou, např. skutečnými gramatickými rozdíly mezi oběma vzorky.
Nemáme-li k dispozici software, který počítá za nás, pak při použití chi-squared testu nejdříve vypočítáme tzv. degree of freedom (d.f.). Vypočítáme jej takto:
(počet sloupců v tabulce frekvencí – 1) / (počet řádků v tabulce frekvencí -1)
(2-1)/(2-1)=1/1=1
Nyní se musíme podívat do tabulek hodnot chi-square a z nich vyčíst pravděpodobnostní hodnotu sloupce. Je-li pravděpodobnostní hodnota blízká nule, pak je silně signifikantní výsledek ve sloupci, je-li blízká 1 je náhodná. V intervalu od nuly do jedné se pohybují signifikantní a nesignifikantní výsledky.
Z našeho výpočtu jsme zjistili, že poměr počtu sloupců zmenšený o jeden a počtu řádků zmenšený o jeden je jedna. Tudíž degree of freedom je roven 1. V tabulce si najdeme, že hodnota pravděpodobnosti pro tuto chi-square hodnotu je 0,0001 (jedna desetitisícina), tj., že je menší než 0,05 (5 setin). Je tedy blíž nule než jedné. Z toho pak můžeme na základě chi-square testu s poměrně vysokým stupněm určitosti vyvodit, že různé výsledky frekvencí nalezené ve srovnávaných textech, jsou odrazem různosti textů a nejsou výsledkem náhody.

Význam kolokací

To že kolokace (souvýskyt) slov má pro lingvistická bádání význam se v lingvistice všeobecně uznává. (Slovníky, frazeologická spojení, ale i v NLP).
V oblasti KL se pro výpočet významnosti kolokací používá dvou technik mi-score (mutual information score) a Z-score.
MI-score je formulka vypůjčená z obl. teoretické computer science a teorie informace. MI-score mezi dvěma slovy (jednotkami) srovnává pravděpodobnost, že dvě jednotky stojí vedle sebe vzhledem k tomu, že k sobě nějakým způsobem patří, nebo zda jsou vedle sebe jen náhodou. Např. nalezneme-li vedle sebe slova riding boots (jezdecké boty) asi sami odhadneme, že stojí vedle sebe proto, že tvoří víceslovné pojmenování, zatímco slova formulka vypůjčená, která jsem užila v předcházející větě, stojí vedle sebe náhodou (formulka vzatá, přejatá, může,….). Čím jsou dvě slova pevněji spjata, tím je mi-score vyšší. Je-li souvýskyt dvou jednotek řídký, nabývá mi-score negativních hodnot. Vyskytnou-li se dvě jednotky vedle sebe náhodou, blíží se mi-score nule.
Podobné výsledky získáme při použití Z-score. Pro každou danou jednotku v textu tento test srovná aktuální frekvenci s frekvencemi souvýskytu s ostatními jednotkami v zadaném okně (např. okně vymezeném tři slova vpravo od dvojice, tři slova vlevo

od dvojice). Čím je Z-score jádrového slova a slov okolních, tím jejich kolokabilita (souvýskyt) pravděpodobnější. Z-score se používá méně než mi-score, ale zmiňuji se o něm proto, že je součástí TACT-concordance package, jednoho s široce dostupných balíčků programového vybavení k počítačovému zpracování korpusů. O mi-score byste měli vědět proto, že v GCQP, korpusovém manažeru používaném v českém prostředí, je součástí vybavení.
Zmíněné techniky mají svůj význam a jsou hojně používané především v lexikografii. Slouží k získání víceslovných jednotek z korpusů. Především se jedná o ty jeednotky, které nejsou známy z tradičních prací pojednávajících o idiomech a frazeologických spojeních. Do těchto prací se kupř. nedostanou víceslovné termíny potřebné při překladech. Druhým případem užití mi-score a Z-score je vytříbení různých významů jednoho slova na základě souvýskytu. Homonymum raketa (tenisová, středního doletu). Ještě zajímavější je to v případě synonym. Tam je oblast využití informací získaných z výpočtu významnosti kolokací použitelná především pro výuku cizinců. Budete se třeba učit anglické slovo strong a powerfull.
Využití je možné i pro strojový překlad (paralelní korpusy).

Zkoumání vztahů mnoha proměnných

Ukázali jsme si různé metody, kterými můžeme podpořit výsledky frekvenční analýzy zjištěními o významnosti získaných výsledků (významné versus náhodné). Pomocí nich můžeme získat dílčí výsledky na jednom korpusu. V KL se ovšem používá rovněž statistických metod, které umožňují srovnat výsledky frekvencí různých proměnných (jednotky, lingvistické jevy,..) v různých vzorcích (textech, korpusech,…) Jedná se o statistické techniky zkoumající multivariantní jevy. Existuje celá řada technik, v KL se používá factor analysis, principal components analysis (analýza hlavních složek), corespondence analysis, multidimensional scaling a cluster analysis.
Nejsem schopna zde uvádět matematické nuance, které stojí za těmito technikami. Jedná se o to, že se používá tabulek kdy v řádcích se uvádějí zkoumaní proměnné a ve sloupcích výsledky frekvence proměnných v různých samples (vzorcích). CROSS-tabulation proměnných a vzorků. Z tabulek se pak počítají matrice korelací odhalující souvislosti mezi výskytem proměnných v různých vzorcích. To je úplně zestručněno podstata faktorové analýzy. Korespondenční analýza pracuje s podobnými principy s tím, že výsledky frekvencí se nanášejí na osy x a y.
O ostatních technikách hovořit nebudu, protože jsem se s nimi ani nesetkala. Nicméně jsem pokládala za dobré se o nich na této přednášce zmínit. Tato přednáška nemůže suplovat úvod do složité matematické statistiky.

Loglinear modely

Jinou technikou sledující vztahy mezi proměnnými v různých vzorcích jsou tzv. loglinearní modely. Tato metodologie nám umožňuje vypočítat, který z výsledků získaných srovnáním frekvenční analýzy formou křížových tabulek je z hlediska statistiky odpovědný za určitý výsledek.
(chi-square test na vyšší úrovni).

Probabilistické modely

Frekvenční data získaná z korpusu se nejčastěji používají pro probabilistické modelování jazyka. Jedná se o metody zpracování PJ (NLP, analýza, generování) na základě výsledků statistického zpracování dat. Tyto techniky se dnes hojně a s nadšením používají, nicméně musím alespoň upozornit na to, že mají své odpůrce v zastáncích systémů budovaných na pravidlech popisujících fungování přirozeného jazyka. Jedním z příkladů použití metod probabilistických modelů je označkování SYN2000, SYN2005.

Shrnutí

V dnešní přednášce jsme si tedy řekli, jak se korpusy, které byly primárně pokládány za zdroj kvantitativních údajů o jazyce mohou díky užití různých metod převzatých z matematické statistiky stát spolehlivějším zdrojem kvantitativních údajů a jak díky zmíněným metodám lze precisně ověřit význam kvantitativních výsledků.