Vědecká práce v lexikologii i lexikografii s ohledem na statistické metody:

In: Klimeš, Lumír: Úvod do vědecké práce v jazykovědné bohemistice (se zvláštním zřetelem k pracím seminárním a diplomovým). Západočeská univerzita v Plzni. Plzeň 2001 (výběr)

Bibliografie:

- definice: srov. SSČ, Akademický slovník cizích slov, Encyklopedický slovník aj. (Klimeš 2001, s. 17n.)

Bibliografická informace: sekundární informace, která slouží k identifikaci dokumentu, jeho obsahu a času... (Klimeš 2001, s. 17-18)
Bibliografická citace: souhrn údajů o citované publikaci nebo její část, umožňující její identifikaci... (dále: ČSN 01 0197, ISO 690; Klimeš 2001, s. 18)
Bibliografický soupis (bibliografie, bibliografický seznam): sekundární dokument obsahující soubor bibliografických záznamů o existujících dokumentech (resp. jejich částech), sestavený podle předem stanovených zásad... (ČSN 01 191; Klimeš 2001, s. 18)

Výklad k české a slovenské bibliografii (Klimeš 2001, s. 18-21); nutná aktualizace textu
K problematice **výpisků**... (Klimeš 2001, s. 30-33)

Kartotéka a klasifikační systém oborů (Klimeš 2001, s. 33-39):

1. různé typy klasifikace:
1.1 podle Bibliografie české lingvistiky: (s. 33-36): lexikální soubory (slovní zásobu, lexikon, jeho původ/ etymologie a vývoj), problematika lexikologie i lexikografie, terminologie, onomastických slovníků apod.
1.2 podle V. Šmilauera (s. 36-38: jazykové vrstvy v širším smyslu, problematika slova, etymologie, slovník)
1.3 metodické přístupy (jazykové vyučování s slovní zásoba; s. 39)
1.4 knihovnický katalog

Základní statistické metody (s. 39n.):

1. ***lexikologie a obecné pojetí lingvistiky*** (39-40):

1.1 kvantitativní lingvistika a frekvenční slovníky
1.2 algebraické

2. ***extenze analyzovaného textu*** (typ vzorku) při lexikální analýze:

2.1 náhodný (aleatorní) výběr položek, stránek apod.:
- počítačově: v nedávné době generátory náhodných čísel, dnes systémové programy
- dříve „ručně“: tabulky náhodných čísel

2.2 počet položek:
a) běžný rozsah: vyšší počet slov než 3000
b) u spec.zaměření (zvl. typy souborů, autorský slovník, tematika aj.): více než 4000 slov (Těšitelová, M.: *Otázky lexikální statistiky*. Praha 1974, s. 18, 21... dále: TOLS)
c) pokud jde o syntaktický rozbor odborného textu, potom se vychází cca z 1000 vět za sebou následujících; větou zde je míněna predikační jednotka nebo i jednočlenná

struktura; bez ohledu na to, zda stojí samostatně – jako tzv. věta jednoduchá – nebo v souvětí: TOLS 29; dále též: Uhlířová, L.: *O délce věty*. SaS 32/1971, s. 232-240)

2.3 konkrétní typy vzorků (Klimeš, s. 40n.):

2.31 ***reprezentativní***: je-li základní soubor vzorkem dokonale zastoupen (reprezentován; ***reprezentace*** ...z lat. *re-praesentatio*, znázornění, zpřítomnění:

a) znázornění, zobrazení, průmět:
- *grafická*
- *vektoru*

b) v matematických popisech lexikologických struktur:
- zobrazení nějaké matematické struktury do algebry matic; např. reprezentace grup

c) zastupování, představování určité společenské skupiny; platí i v rovině komunikace)

2.311 U vzorku získaného metodou náhodného výběru:každý prvek (jednotka) základního souboru se vyznačuje stejnou pravděpodobností, že se stane prvkem (jednotkou) reprezentativního vzorku (prvky musí být vzájemně srovnatelné)

2.32 ***standardní***: pokud je vzorek upraven tak, aby byl srovnatelný s jiným vzorkem – vzhledem k témuž ukazateli (sledujeme stejný problém či typ problému na větším množství vzorků)

3. ***některé pojmy z matematické statistiky*** (41n.):

3.1 ***aritmetický průměr*** = součet hodnot vydělený počtem členů souboru (x, μ):

a) sčítají se i hodnoty stejné velikosti, nula apod. (nelze vynechávat při stanovení hodnoty dělitele)
b) hodnoty průměru a procentuální vyjádření mají menší význam pro poznání struktury souboru, než se obvykle soudí

Problémy:

- nebezpečí rozložení četnosti výskytu jednotlivých hodnot v asymetrickém modelu:
  - dochází k posunu k pravostranné či k levostranné pozici
  - lepší je volit průměr *harmonický* nebo *geometrický*, popř. vycházet z tzv. *mediánu* (tj. z prostřední hodnoty souboru; srov. dále sub 3.2)
- nesnáze s určením prvků aritmetického souboru (nelze systematicky určovat):
  - podle extrémních (krajních), maximálních či minimálních hodnot
  - míru stejnorodosti (homogenity:jak „daleko“ jsou jednotlivé hodnoty od aritmetického průměru)
  - formální hodnotu (tvar) souboru (hodnoty větší či menší než aritmetický průměr)
  - frekvenci (nejčastější opakování) hodnot, popř. skupin hodnot... Pokud jsou rozdíly v počtu členů souborů velké, procentuální vyjádření vede k nesprávným závěrům. Potom je nutné údaje ověřit *statistickým testem významnosti* (srov. např. Fabián, V.: *Základní statistické metody*. Praha 1963... an.)
- nelze jednoduše sčítat procentuelní zastoupení jednotlivých položek a poté dělit počtem prvků (počítáme mnohdy procenta z různých základů!)

- pokud se aritmetické průměry dvou hodnotových škál shodují, nelze z toho odvodit blízkost sledovaných postupů ani v případě, že mají obě škály stejný počet členů (položek)!
- jestliže je rozložení četností dvou- a vícevrcholové, není třeba počítat aritmetické průměry. Stačí porovnat *mediány*...
- porovnávání průměrů 2 a více souborů je statisticky průkazné (významné), není-li mezi počtem jejich členů statisticky významný rozdíl (mezi hodnotami jejich „rozptylů“)
- nelze jednoduše počítat „průměry z průměrů“ (pokud neobsahují všechny započítané průměry stejný počet položek)

3.2 ***medián***:

3.21 určení:

- prostřední hodnota v souboru seřazeném podle velikosti (vzestupně, někdy sestupně)
- prostřední hodnotou se míní prostřední položka, nikoliv průměr nejnižší a nejvyšší hodnoty souboru
- jestliže má soubor sudý počet položek, vybereme po jedné položce umístěné nalevo a napravo od středu a vypočítáme jejich aritmetický průměr. To je v tomto případě *medián*.

3.22 význam:

- umožňuje získat představu o tom, jak daleko je aritmetický průměr od středu souboru
- odhadneme tak směr asymetrie souboru (zejména při extrémním vybočení některých hodnot souboru z běžné normy)
- podklad pro využití *mediánového testu*

3.3 ***modus***:

3.31určení:

- hodnota, popř. více hodnot, která/ které se v souboru vyskytuje/-í nejčastěji
- některé soubory nemají žádný *modus*

3.32 význam:

- důležitý je vztah hodnoty *modu* a *aritmetického průměru*
- určení polohy *modu* ve vztahu k *mediánu* (totožnost, poloha vlevo – vpravo od *mediánu*...)
- určení nejčastější hodnoty vzorku (k pochopení struktury souboru)

3.33 *modus* uvádíme, obsahuje-li min. ¼ až ⅓ všech členů souboru

3.4 ***variační rozpětí***:

- také ***variační šíře*** (***R***): rozdíl mezi největší a nejmenší hodnotou
- význam:
- umožňuje uvážit rozdíl hodnot, lépe popsat míru nahromadění hodnot kolem průměru (ale pouze z hlediska hodnot extrémních!)

3.5 ***variance, statistický rozptyl***

3.51 podává informace o (způsobu) rozložení hodnot kolem aritmetického průměru

3.52 dvojí způsob výpočtu:

a) pro malé soubory (od 3 do 31 prvků): $s^2 = \frac{\sum x_i^2}{n-1} - \bar{x}^2$

b) pro velké soubory (od 32 prvků): $s^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$

3.6 ***směrodatná odchylka***:

- kladně pojímaná druhá odmocnina ze statistického rozptylu: $s = +\sqrt{s^2}$
- orientace i v rámci extrémních (krajních) hodnot souborů
- je nutné, aby *rozložení četnosti* (srov. sub 3.0) jednotlivých souborů bylo *normální*

3.61 výpočet základního rozmezí hodnot v intervalu:

a) $x \pm s$ (obvykle 68 %všech hodnot souboru): hodnoty blížící se průměru

b) $x \pm 2s$ (obvykle 95 %všech hodnot souboru): ani tento interval nezahrnuje extrémní hodnoty

3.62 míra stejnorodosti, homogenity souboru:

- velikost intervalu závisí na velikosti vypočtené směrodatné odchylky *s*: čím větší, tím větší rozpětí – tím větší rozptýlení hodnot kolem aritmetického průměru
- dva soubory mohou mít stejný *aritmetický průměr* i *variační rozpětí* – a přesto se od sebe mohou značně lišit strukturou
- směrodatnou odchylku vyjadřujeme ve stejných jednotkách jako aritmetický průměr i jako jednotlivé členy souboru (jinak by ji ani nebylo možné vypočítat!)

3.7 ***variační koeficient*** (V_k):

- při porovnání dvou souborů můžeme zjistit značný rozdíl v *aritmetickém průměru* i ve *směrodatné odchylce*...
- ke zjištění rozdílů mezi homogenností souborů“procentuální výpočet poměru *směrodatné odchylky* vůči *aritmetickému průměru*... (kolik procent tvoří *odchylka* v každém souboru z *průměru*): $V_k = \frac{s}{\bar{x}} 100 = \frac{100s}{\bar{x}}$

3.8 ***střední chyba*** (s_E):

- ukazuje, jaké chyby se dopustíme, jestliže budeme statistický soubor charakterizovat *aritmetickým průměrem*
- vzorec k výpočtu: $s_E = \frac{s}{\sqrt{n}}$
- velmi malá (většinou zanedbatelná), pokud nepřesáhne 5 % aritmetického průměru

3.9 ***rozložení četnosti***:

- při porovnání hodnot aritmetického průměru, variačního koeficientu, statistického rozptylu, směrodatné odchylky a střední chyby (přihlížíme přitom k hodnotě modu a mediánu)
- na základě výpočtu lze zjistit, zda se rozložení četnosti zkoumaného souboru odchyluje od rozdělení normálního do té míry, že zkoumané rozdělení již za normální považovat nelze...
- podrobněji: Klimeš 2001, s. 46 – 51 (včetně grafického vyjádření)

3.0 ***hladina významnosti***:

- sledujeme míru pravděpodobnosti (např. při potvrzení či vyvrácení nějaké hypotézy\)
- pokud hladina významnosti dostoupí 95 % (a více), je rozdíl velice významný!
- jedná se tu o stanovení statisticky významného (signifikantního) rozdílu