MATEMATICKÁ LINGVISTIKA

RNDr. Blanka Sedlačíková, Ph.D.

MATEMATICKÁ LINGVISTIKA

1. hraniční disciplína stojící mezi matematikou a lingvistikou

2. část lingvistiky využívající matematických metod

CÍL

- exaktní popis přirozeného jazyka opřený o matematické metody

VZNIK

- 50. léta 20. století (VIII. mezinárodní lingvistický kongres v Oslo v roce 1957)

DĚLENÍ MATEMATICKÉ LINGVISTIKY

- lingvistika kvantitativní
- lingvistika algebraická
- lingvistika počítačová

VYUŽITÍ MATEMATICKÝCH METOD V LINGVISTICE

1. kvantitativní lingvistika – teorie pravděpodobnosti, matematická statistika, teorie informace, matematická analýza aj.

2. algebraická lingvistika – algebra, teorie množin, teorie grafů, kombinatorika, logika aj.

3. počítačová lingvistika – praktická aplikace dvou předchozích

1. LINGVISTIKA KVANTITATIVNÍ

- vznik na konci 50. let 20. stol.
- ta část matematické lingvistiky, která využívá kvantitativních metod
- podle převládající metody bývá někdy používáno označení „lingvistika statistická"

- **starověk** – Hindové (Rgvédy – hymnické texty bráhmanů – 2.tis. př. n.l.)
- **středověk** – mystika, hříčky
  - kabala
  - obrazové básně
  - steganografie, tj. tajná písma (→ teorie kódování, → kombinatorika – J. Wilkins, G. W. Leibniz – 1666 *Dissertacio de arte combinatoria*)
  - spekulativní gramatiky (vedle popisu jazyka využívají postupy logické a filozofické)
  - filozofické gramatiky (→ generativní transformační mluvnice N. Chomského)

Kabala

- proud hebrejského (židovského) mysticismu
- stvoření světa považuje za jazykový jev
- matematické postupy uplatňované v kabale:

→ Notarikon – metoda akrostichu (počáteční písmena některých slov dávají slovo jiné) – *ars notoria*

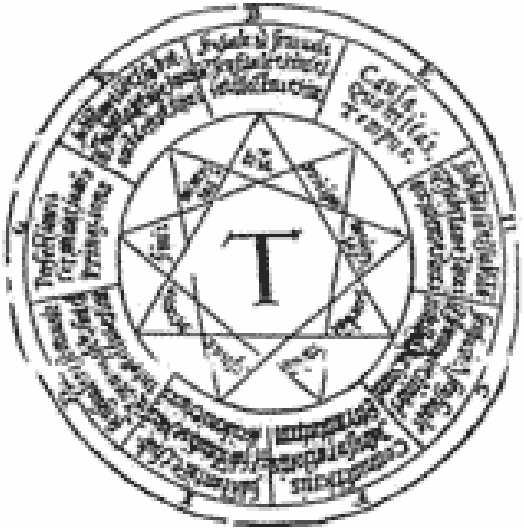
→ Gematrie – pracuje s číselnou hodnotou písmen (Př.: JHVH – 72; had i mesiáš – 358)

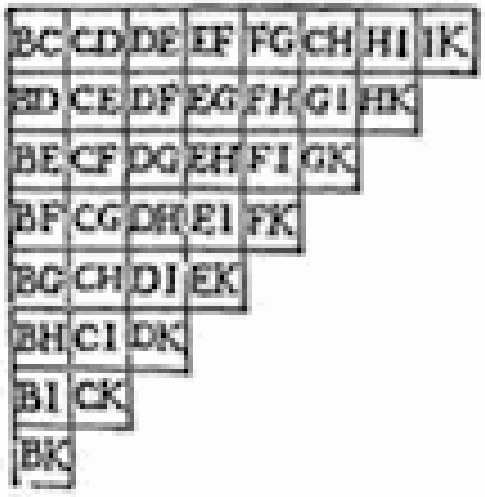
→ Temura – umění permutace, tj. záměny písmen, umění anagramu

Univerzální jazyk

- Raimundus Lullus (1232 – 1316)

  → kombinatorika (G. W. Leibniz)





POJEM „frekvence“

= počet (četnost) určitého jevu v celku

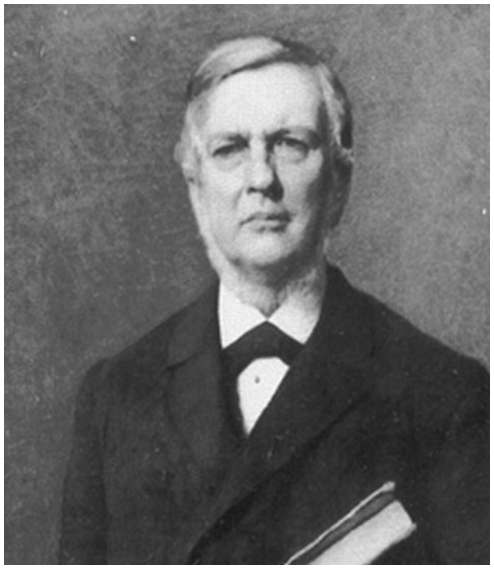
- tiskaři
- těsnopis
- pedagogická praxe
- kódování (Morseova abeceda)
- klaviatura psacího stroje
- šifrování a dešifrování textu
- frekvenční a konkordanční slovníky

VIKTOR JAKOVLEVIČ BUNJAKOVSKIJ

- 1847 – časopis *Sovremennik* – možnosti využívání matematických metod v ling-vistice

О ВОЗМОЖНОСТИ

ВВЕДЕНІЯ ОПРЕДѢЛИТЕЛЬНЫХЪ МѢРЪ ДОВѢРІЯ КЪ РЕЗУЛЬТАТАМЪ НѢКОТОРЫХЪ НАУКЪ НАБЛЮДАТЕЛЬНЫХЪ, И ПРЕИМУЩЕСТВЕННО СТАТИСТИКИ.



Ernst Förstemann (1822–1906)

August Schleicher (1821–1868)

- Förstemann (1846, 1852) – frekvence hlásek a písmen v němčině, řečtině, latině a gótštině
- Schleicher (1852) – frekvence písmen a hlásek ve staré církevní slovanštině (doplněk k pracem E. Förstemanna)

ANDREJ ANDREJEVIČ MARKOV (1856 – 1922)

- 1913 – *Primer statističeskogo issledovanija nad tekstom „Jevgenija Onegina“, illjustrirujuščij svjaz ispytanij v cep* – hlásková statistika veršovaného románu Evžen Oněgin, na který aplikoval svou teorii markovských řetězců

NÁPODOBA TEXTU PODLE TEORIE PRAVDĚPODOBNOSTI

1) Za předpokladu, že všechna písmena v textu mají stejnou frekvenci:

čeština	*ďj mrgučxýďyaýweaožá*
angličtina	*xfoml rxkhrjff juj zlpwcfwkkcyj*
němčina	*aiobnin tarsfneoulpiitdregedcoads*

2) S přihlédnutím k relativní frekvenci jednotlivých písmen:

čeština	*žia ep atndi zéuořmp*
angličtina	*ocro hli rgwr nmielwis eu ll*
němčina	*er agepterprteiningeit gerelen re*

3) S přihlédnutím k relativní frekvenci dvojic písmen:

čeština	*lí di oneprá sguluvicechupsv*
angličtina	*on ie antsoutinys are t inctore*
němčina	*billunten zugen hin se sch wel*

4) S přihlédnutím k relativní frekvenci trojic písmen:

čeština	*dves a vaše miléklár*
angličtina	*in no ist lat whey cratict froure*
němčina	*eist des nich in den plassen kann*

Prvních deset slov s nejvyšší frekvencí v různých jazycích

Pořadí	Čeština	Slovenština	Angličtina	Ruština
1.	a	a	the	v(o)
2.	být	byť	of	i
3.	ten	v	and	ne
4.	v(e)	na	a	na
5.	on	sa	in	ja
6.	na	ten	that	byť
7.	že	on	is	čto
8.	s(e)	že	was	on
9.	z(e)	z	he	s(o)
10.	který	ako	for	a

Slovní zásoba některých spisovatelů

- France – 9 000 slov
- Homér – 9 000 slov
- J. W. Goethe – 20 000 slov
- A. S. Puškin – 21 200 slov
- W. Shakespeare – 24 000 slov

Slovník některých děl

- K. Čapek: Obyčejný život – 5 539 slov
- K. Čapek: Život a dílo skladatele Foltýna – 4 145 slov
- M. Pujmanová: Předtucha – 4 858 slov
- Fr. Halas: Ladění – 2 078 slov
- J. Hora: Kniha domova – 2 961 slov

Průměrný slovník dětí

- 1 rok – 10 slov
- 2 roky – 300 slov
- 3 roky – 900 slov
- 4 roky – 1.650 slov
- 5 let – 2.500 slov
- 6 let – 3.500 slov
- 14 let – 9.000 – 19.500 slov

Slovník dospělých podle profesí (Ogden)

- farmář – 300 slov
- japonský diplomat – 7.000 slov
- univerzitní student – 12.000 slov
- James Joyce – 250.000 slov

K běžnému dorozumění v cizím jazyce je podle Ogdena třeba znát asi 850 slov (600 podstatných jmen, 150 přídavných jmen, 100 sloves). K sledování odborné literatury potom ještě dalších 150 (100 termínů obecně vědeckých a 50 termínů z daného oboru) – celkem tedy 1.000 slov.

GEORGE KINGSLEY ZIPF

(1902 – 1950)

První Zipfův zákon

$r \cdot f = k$

(součin frekvence slova a jeho ranku je konstantní)

Druhý Zipfův zákon

$a \cdot b^2 = k$

(počet slov o jisté frekvenci krát frekvence na druhou je konstantní)

Třetí Zipfův zákon:

$$\frac{rr}{\sqrt{f}} =$$

(slova s vysokou frekvencí mají zpravidla větší počet významů)

GLOTTOCHRONOLOGIE

- též lexikostatistika
- lexikologická metoda, která pomocí statistiky zjišťuje dobu vzniku jazyka, respektive různých jazyků
- vznik 50. léta 20. století (M. Swadesh)
- inspirace tzv. radiokarbonovou metodou (americký chemik Willard Frank Libby)

PRINCIP: 1) slovníkové jádro

2) rychlost změn v jádru

Časová hloubka se vyjádří pomocí vzorce

$$i_{(t)} = \frac{lnC}{2lnr}$$

$i_{(t)}$ … časová hloubka (uběhlý čas);

C … procento společně zděděných párů slov z počtu všech slov v slovníkovém jádru obou zkoumaných jazyků;

r … index rychlosti mizení slov z jádra (r=0,86)

TEORIE INFORMACE

Matematická disciplína zabývající se přenosem, kódováním a měřením informace (C. E. Shannon, W. Weawer)

Entropie (H)

- Míra neurčitosti pokusu (průměrné množství informace obsažené v jednom výsledku příslušného pokusu)

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i$$

N … počet prvků v množině,

p_i … pravděpodobnost výskytu i-tého prvku pro i = 1, 2, …, N

Redundance

Udává procento nadbytečných jednotek sdělení o entropii H_n (entropii n-tého řádu)

$$R_n = 1 - \frac{H_n}{H_0}$$

Bit

Jednotka množství informace (je daná abecedou o jednom prvku a dvou stavech)

Šum

Označení jakékoli poruchy přenosu informace

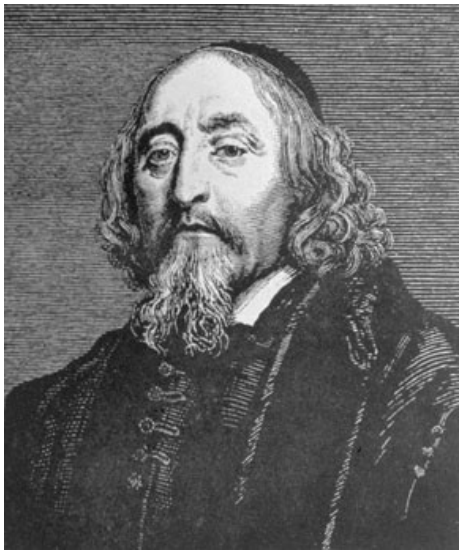
Český přínos matematické lingvistice

- J. A. Komenský (1631)
- časopis Krok (1831)
  - hlásková statistika
  - srovnání slovní zásoby různých jazyků
- A. Seydler (1886)
- F. Wolf (1929)
- pražská lingvistická škola

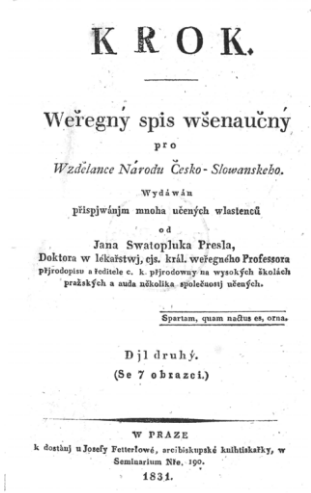
Jan Amos Komenský


J. A. Comenii
JANUA LINGUARUM
reserata.
ANNO M. DC. LXVI
J. A. Comenii
Ontsloote Deure der Talen.

- 1631 – *Janua linguarum reserata* (*Brána jazyků otevřená*) – ekonomické rozvíjení slovní zásoby žáků
- hledání univerzálního jazyka
- moderní myšlenky – strukturalismus (rozlišování aktivní a pasivní slovní zásoby, jednota formy a obsahu aj.)



Časopis Krok a první užití matematiky v lingvistice u nás



KROK.

Weřegný spis wšenaučný
pro
Wzdělance Národu Česko-Slowanskeho.
Wydáwán
přispjwánjm mnoha učených wlastenců
od
Jana Swatopluka Presla,
Doktora w lékařstwj, cjs. král. weřegnho Professora
přjrodopisu a ředitele c. k. přjrodowny na wysokých školách
pražských a auda několika společnostj učených.

Spartam, quam nactus es, orna.

Djl druhý.
(Se 7 obrazci.)

W PRAZE
k dostánj u Josefy Fotterlowé, arcibiskupské knihtiskařky, w Seminarium Nro. 190.
1831.

- 1831 – **časopis Krok** (1821–1840) – hlásková statistika češtiny, němčiny a italštiny (*Srovnání spoluhlásek a samohlásek v češtině, vlaštině a němčině*)
- srovnání slovní zásoby různých jazyků (*Český jazyk s jinými jazyky z ohledu bohatosti slov porovnaný*) a zastoupení jednotlivých slovních druhů (s výjimkou španělštiny)

Fotokopie hláskové statistiky v časopise Krok (1831)

Srownánj spoluhlásek a samohlásek w češtině, wlaštině a němčině.

Wyčjtá se wůbec češtině, že pro náramné množstwj spoluhlásek gest twrdá a těžká k wyslowenj. K wywrácenj této wýčitky poslaužj následugjcj porownánj.

W 1000 pjsmenách

	českých		wlaských		německých	
samohlasek	480,0		512,0		326,0	
a . .	120,0	—	130,0	—	80,0	(aa, ah, au, ei aei, ei, eu,ey.)
e . .	110,0	—	135,0	—	77,5	(ä, ee, oe, eh,*)
i . .	90,0	—	110,0	—	127,5	(ie,ü,y,i w ai a d.)
		některé ğ y,j.				
o . .	100,0	—	92,0	—	20,0	(oo, oh)
u . .	60,0	—	45,0	—	22,0	
spoluhlasek . . .	520,0	—	488,0	—	674,0	
b . .	10,0	—	17,5	—	38,0	
c . .	10,0	—	2,5(z)	—	22,5	(z)
č . .	4,0	—	5,0	—	00,0	
d . .	50,0	—	25,0	—	66,0	
f . .	00,0	—	5,0	—	18,0	(v)
ğ . .	00,0	—	7,0(gh)	—	25,5	
h . .	2,5	—	2,5	—	50,0	
ch . .	17,5	—	0,0	—	12,0	
j (g) .	10,0	—	2,5	—	2,5	
k . .	42.5	—	45,0 (c před a, o,u,kw qu)		22,5	
l . .	40,0	—	70,0(ll)	—	20,0	(ll)

*) *e* němá gsau právem wynechána, poněwadž se newyslowugj.

	českých		wlaských		německých	
m . .	37,5	—	22,0(mm)	—	29,0	(mm)
n . .	64,0	—	90,0(nn)	—	127,5	(nn)
p . .	22,5	—	27,0	—	5,0	
q . .	wiz k a w					
r . .	25,0	—	43,0	—	70,0	
ř . .	2,5	—	00,0	—	00,0	
s . .	27,5	—	37,0	—	55,0	
š . .	30,0	—	7,5(sci)	—	36,5	(sch)
t . .	57,0	—	55,0	—	52,0	
w . .	27,5	—	15,0(v, w w qu)		22,0	
x . .						
z . .	30,0	—	wiz s	—	wiz s	
ž . .	10,0	—	2,5 (gi)	—	00,0	

Na gednu samohlásku přigde w češtině spoluhlásek 1,083, we wlaštině 0,952, w němčině 2,067. Na gednu spoluhlásku přigde w češtině samohlásek 0,923, we wlaštině 1,049, w němčině 0,483.

Odhad slovní zásoby a počtu slovních druhů v některých jazycích (Krok 1831)

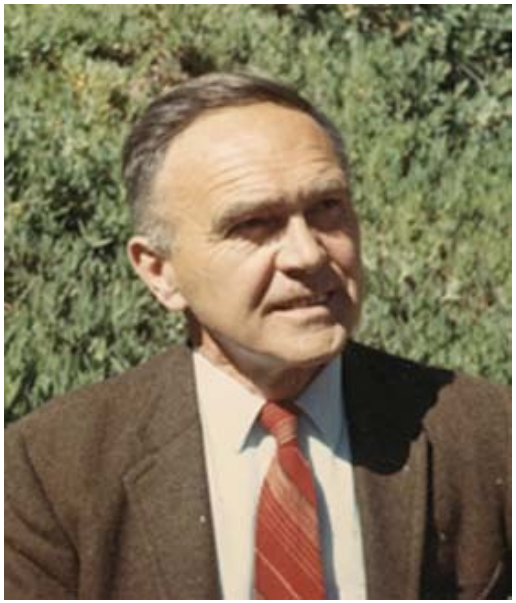
Jazyk	Počet slov
Němčina	29.000
Španělština	30.000
Francouzština	32.000
Italština	35.000
Angličtina	37.000
Čeština	57.000

Slovní druh	Němčina	Francouz-ština	Italština	Angličtina	Čeština
podst. jména	21 395	19 706	13 030	16 127	21 276
slovesa	2 655	5 657	7 877	10 141	13 184
přídavná jména	2 475	4 803	5 843	8 444	14 184
příslovce aj.	2 475	1 834	8 250	2 888	14 356
celkem	29 000	32 000	35 000	37 000	57 000

AUGUSTIN SEYDLER

1886 – článek *Počet pravděpodobnosti v přítomném sporu* v časopise Athenaeum (tzv. rukopisný spor)

FRANTIŠEK WOLF



Wolf, F.: *Použití počtu pravděpodobnosti k identifikaci textu*. Inaugurace rektorů v Brně 1928/29, 1929/30, s. 99–105 (využití počtu pravděpodobnosti k řešení jazykovědné povahy).

- narodil se 30. 11. 1904 v Prostějově
- studium matematiky a fyziky na brněnské univerzitě (profesoři B. Hostinský a E. Čech)
- působení v Cambridgi
- 1938 – habilitace (u prof. Jarníka na PřF UK)
- stipendium ve Švédsku (Mittag-Leflerův ústav, u prof. Carlemanna) – od 1. 12. 1938
- 1941 – emigrace do Ameriky
- od 1942 – University of California v Berkeley
- Dílo: konvergence obecných trigonometrických řad, funkcionální analýza (teorie perturbací), teorie singulárních hraničních problémů v parciálních diferenciálních rovnicích
- zemřel 12. srpna 1989

2. LINGVISTIKA ALGEBRAICKÁ

- formuje se od konce 50. let 20. století
- ta část matematické lingvistiky, která využívá nekvantitativní matematické metody (zejména matematickou logiku, teorii množin, teorii grafů, algebru)
- N. Chomsky (generativní gramatika)
- Y. Bar-Hillel (kategoriální gramatika)
- S. K. Šaumjan
- S. Marcus aj.

GENERATIVNÍ GRAMATIKA (Noam Chomsky)

1. Věta → Podm + Přís
2. Podm → Subst
3. Subst → Adj + Subst
4. Přís → Slov
5. Slov → Adv + Slov
6. Adj → Adv + Adj
7. Subst → *muž*
   žena
   lavice
8. Adv → *hodně*
   málo
   pěkně
9. Slov → *píše*
   jí
   váží
10. Adj → Km Konc
11. Km → *mlad*
    velk
    větš
12. mlad Konc + *muž* → *mlad*-ý + *muž*
13. mlad Konc + žena → *mlad*-á + *žena*
14. velk Konc + *muž* → *velk*-ý + *muž*
15. velk Konc + *lavice* → *velk*-á + *lavice*
16. velk Konc + *žena* → *velk*-á + *žena*
17. větš Konc → *větš*-í

Velká lavice hodně váží.

Velký muž hodně váží.

Mladá žena málo jí.

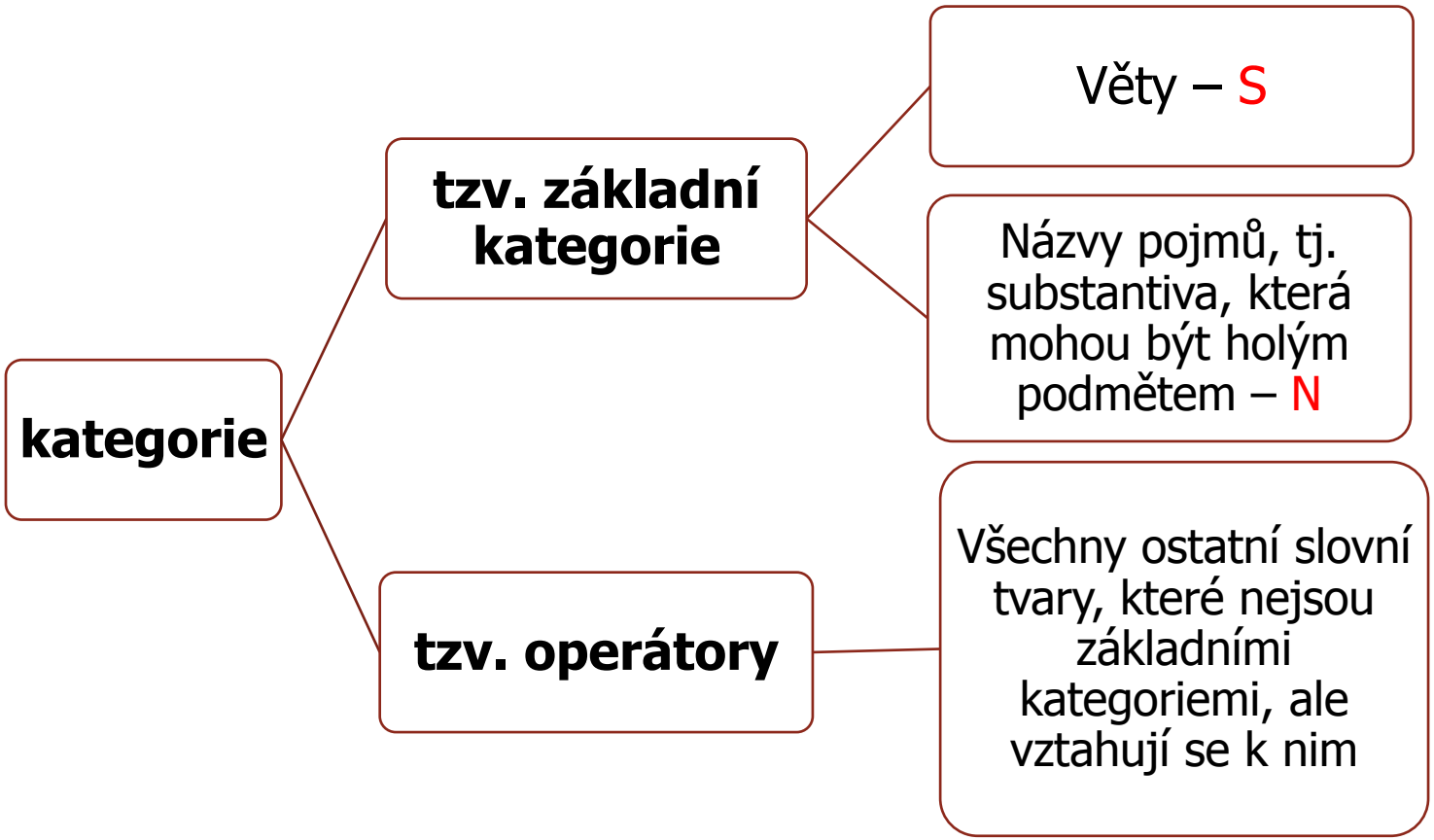
Velká lavice pěkně jí.

Mladá mladá žena jí.

KATEGORIÁLNÍ GRAMATIKA (Y. Bar-Hillel)

1) Každou větu lze přepsat jako posloupnost symbolů.

2)



3) Vedle jednoduchých kategorií S a N rozlišuje složené kategorie:

N/N, (N/N)/(N/N)

4) Zavádí symboly „/ “ (nad) a „\“ (pod)

- „N/N“ (N nad N) – odpovídá např. adjektivu, které ve větě předchází před příslušným substantivem, na němž je závislé

- „N\S“ (N pod S) – odpovídá zpravidla slovesu, jehož tvar je řízen podmětem, který před slovesem předchází

5) Řetězy symbolů lze krátit podobně jako zlomky; rozlišuje se:

- krácení zprava

$$N/N,\ N \rightarrow N$$

(podobně jako $x/8 \cdot 8 = x$)

- krácení zleva

$$N,\ N\backslash S \rightarrow S$$

(podobně jako $8 \cdot 1\backslash 8 = 1$)

6) Každou větu jazyka lze nahradit řetězem symbolů – často je možností více. Na každý z nich lze aplikovat podle potřeby krácení zleva a zprava. Získáme-li alespoň v jednom případě jako výsledek jednoduchý symbol, pak se jedná o **gramaticky správnou větu daného jazyka**.

Dobrý vysokoškolský student samostatně přemýšlí.

N/N, N/N, N, (N\S)/(N\S), N\S

N/N, N/N, N, N\S
(dobrý vysokoškolský student přemýšlí)

N/N, N, N\S
(dobrý student přemýšlí)

N, N\S
(student přemýšlí)

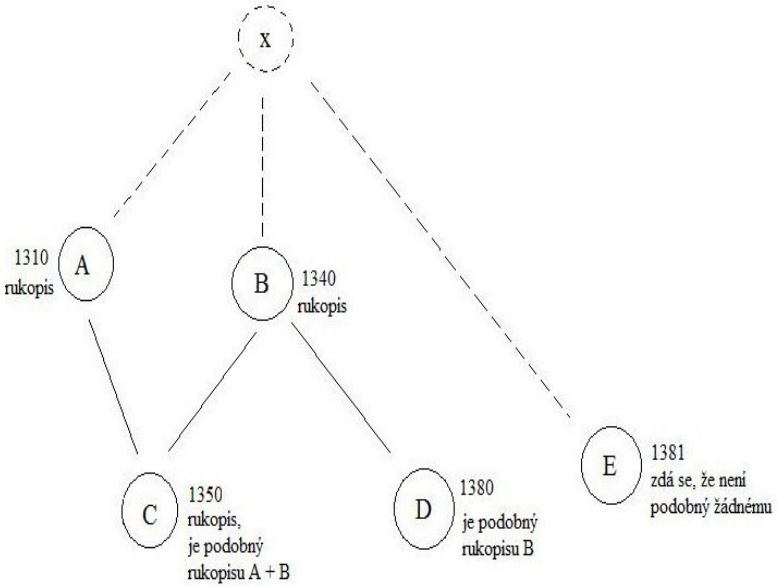
S
(přemýšlí)

TEXTOLOGIE

- též textová kritika
- společenskovědní obor stojící na pomezí literární vědy a jazykovědy, který podrobuje text všestranné analýze

Využití matematických metod

1) určování autorství
2) stylometrická metoda
3) sestavování stemmatu
   - *metoda společných chyb*
   - *taxonomická metoda*
   - *metoda rozkladu množiny*



3. POČÍTAČOVÁ LINGVISTIKA (STROJOVÁ, KOMPUTAČNÍ)

- od konce 50. let 20. století
- praktická aplikace lingvistiky kvantitativní a algebraické (v podstatě jde o počítačové zpracování jazyka)

Problematika

- uchovávání a vyhledávání informací
- strojový překlad
- korpusová lingvistika aj.

KORPUSOVÁ LINGVISTIKA

- část počítačové lingvistiky, která se zabývá tvorbou a využitím jazykových korpusů

KORPUS - rozsáhlý, vnitřně uspořádaný a ucelený soubor jazykových dat, která jsou elektronicky uložena, zpracována a přístupna

POŽADAVKY NA KORPUS

- rozsáhlý
- variabilita
- autenticita
- aktualizace

VYUŽITÍ KORPUSŮ

- lingvistické účely (tvorba elektronických slovníků, frekvenční a statistické studie, výzkum slovní zásoby...)
- sociologie, psychologie
- žurnalistika
- počítačová lingvistika – matematici a programátoři (pro ověřování nových automatických nástrojů k výzkumu jazyka)

ČESKÝ NÁRODNÍ KORPUS

(http://ucnk.ff.cuni.cz)

- 1988 – Iniciativní skupina pro přípravu počítačových korpusů, textů a slovníků
- 1991 – Skupina pro počítačový fond češtiny (odborníci z ÚJČ UK v Praze, MU v Brně, UP v Olomouci)
- 1992 – Český národní korpus jako grantový projekt GA ČR Čeština ve věku počítačů (ÚČNK FF UK – F. Čermák, Ústav formální a aplikované lingvistiky na MFF UK – E. Hajičová, Ústav formální a teoretické lingvistiky FF UK – P. Sgall, V. Petkevič, Katedra českého jazyka FF UK – K. Kučera, ÚČJ FF MU – K. Osolsobě, Katedra informačních technologií FI MU – K. Pala, ÚČJ AV ČR – J. Králík)
- 1994 Ústav Českého národního korpusu na FF UK

Zahraniční korpusy:

- Brown Corpus of Written American English (první počítačový korpus, W. N. Francis a H. Kučera, 1961 – 1964, cca 1 milion slov)
- Bank of English (jeden z nejznámějších, v roce 2012 cca 650 milionů slov)
- British National Corpus (1991, obsahuje 100 milionů slov – 90 % je jazyka psaného a zbytek tvoří jazyk mluvený).
- Helsinki Corpus (diachronní korpus angličtiny)

Studium matem. lingvistiky v ČR

- doktorský stud. program *Matematická lingvistika* (Ústav teoretické a komputační lingvistiky FF UK, Ústav Českého národního korpusu FF UK, Ústav formální a aplikované lingvistiky, MFF UK)
- magisterské studium *Počítačová a formální lingvistika* (MFF UK)
- Od 2010 bakalářké a magisterské studium *Český jazyk se specializací počítačová lingvistika* (ÚČJ FF MU, Centrum počítačové lingvistiky při ÚČJ MU, Katedra informačních technologií FI MU, Centrum zpracování přirozeného jazyka FI MU)

POČET PRACÍ Z MATEMATICKÉ LINGVISTIKY PUBLIKOVANÝCH V ČESKOSLOVENSKU V LETECH 1962–1965

Rok	Počet prací
1962	107
1963	183
1964	143 *(+104)*
1965	210

Zdroj:

Ludvíková Marie: *Bibliografie kvantitativní lingvistiky*. NŘ 50, 1967.

LITERATURA

- Čermák, Fr., Blatná R.: *Manuál lexikografie*. H&H, Jinočany 1995.
- Černý, J.: *Dějiny lingvistiky*. Votobia, Olomouc 1996.
- Sgall,P. a kol.: *Cesty moderní jazykovědy*. Orbis, Praha 1964
- Sedlačíková, B.: *Historie matematické lingvistiky*. Dizertační práce, Brno 2010.
- Sedlačíková, B.: *Matematická lingvistika*. Učitel matematiky 10, 2001/2002, str. 30-36, 80-88, 174-181, 226-234.
- Těšitelová, M.: *Kvantitativní lingvistika*. SPN, Praha 1987.
- Těšitelová, M. a kol.: *O češtině v číslech*. Academia, Praha 1987.
- Vašák, P.: *Matematika, exaktnost a literatura*. Československý spisovatel, Praha 1986.

Děkuji za pozornost!