

# PLIN041 Vývoj počítačové lingvistiky

## Kvantitativní lingvistika I

Mgr. Dana Hlaváčková, Ph.D.

# Formality

- prezentace v Učebních materiálech IS
- další materiály v ELFu
- závěrečný test

# Pár poznámek na úvod

- přibližně 100 let vývoje oborů, které jsou dnes součástí počítačové lingvistiky
- USA, Evropa, část Asie (Sovětský svaz), univerzity, armádní a vládní organizace
- polovina z této doby – práce bez počítačů
- dvě světové války, studená válka, ideologie
- nejde o jednotný obor
- lingvistika spojená s přírodními a technickými obory
- osobnosti s širokým záběrem vědních oborů (často nejde o lingvisty)
- během vývoje neustálený název
  - matematická lingvistika, jazykové inženýrství
- dnes Počítačová lingvistika, Computational Linguistics
- zdroje informací – knihy, články, vzpomínky studentů, nekrology
- <https://docs.google.com/document/d/1pCCpiN8EJcD15pLhd2oq1PE9xzJSQUULVjzflwu32a1/edit?usp=sharing>

# Osnova kurzu (90. léta 19. st.–90. léta 20. st.)

1. Počátky matematického přístupu k jazyku
2. Frekvenční a statistické studie, kvantitativní lingvistika
3. Teorie komunikace a informace. Algebraická lingvistika
4. Strojový překlad. První počítače
5. Korpusová lingvistika a počítačová lexikografie
6. Počítačová lingvistika v ČR

# Počátky matematického přístupu k jazyku

*2. pol. 19. st. – 60. léta 20. st.*

# Počátky (2. pol. 19. st. – 60. léta 20. st.)

- historicko-srovnávací jazykověda (19. st.)
- mladogramatikové (konec 19. st.)
- experimentální fonetika (rozvoj přírodních věd, přístrojové vybavení)
- nové směry počátku 20. st., **vliv přírodních věd**
- Ferdinand de Saussure
- strukturalismus

# Počátky (2. pol. 19. st. – 60. léta 20. st.)

- 19./20. st. – pronikání statistických metod do lingvistiky
- centrem pozornosti – **frekvence** (tiskaři, stenografové, Morseova abeceda)
- americký lingvista (gramatika sanskrtu) **William D. Whitney** (1827–1894) – frekvence anglických hlásek
- německý stenograf **F. W. Käding** – *Slovník četnosti výskytu německého jazyka*, *Häufigkeitwörterbuch der deutschen Sprache*, 1897–98, cca 11 mil. slov (právní a obchodní texty), první frekvenční slovník, frekvence slov, slabik a písmen, 320 výrazů pokrývá  $\frac{3}{4}$  textu

# Počátky (2. pol. 19. st. – 60. léta 20. st.)

- ruský matematik **Andrej Andrejevič Markov** (1856–1922) – *Příklad statistického výzkumu textu Evžena Oněgina...*, 1913
- **Markovův proces** – v procesu mluvení k již vysloveným jednotkám přibývají další podle pravidel jejich relativní frekvence
- teorie pravděpodobnosti, teorie informace
- [**Pavel Novák** – *Teorie informace a lingvistika* (Cesty moderní jazykovědy, 1964), relativní četnost písmen + pravděpodobnost výskytu písmena v závislosti na předchozím písmenu]



# Počátky (2. pol. 19. st. – 60. léta 20. st.)

- americký lingvista německého původu **George Kingsley Zipf**, relativní frekvence hlásek, princip nejmenšího úsilí (20./30. léta 20. st.)
- **1. Zipfův zákon**  
 $r \cdot f = k$ , čím je rank slova nižší, tím je jeho frekvence vyšší (Těšitelová – platí pro střední část frekvenčního slovníku)
- **2. Zipfův zákon**  
 $a \cdot b^2 = k$ , čím je frekvence nižší, tím více slov tuto frekvenci má (a – počet slov, b – sdílená frekvence)
- **3. Zipfův zákon**  
počet různých významů (polysémie) je vyšší u slov s vyšší frekvencí (krátká slova)

# Počátky (2. pol. 19. st. – 60. léta 20. st.)

- po 2. sv. v. vznik pomezních disciplín (matematická lingvistika, sociolingvistika, psycholingvistika, etnolingvistika atd.)
- přelom 50. a 60. let, **1957 VIII. mezinárodní lingvistický kongres v Oslo**
- **Joshua Whatmough – Mathematical Linguistics**
- **matematická lingvistika** – využívání metod přírodních věd (statistika, algebraické metody), seminář na Harvardu
- **kvantitativní** (statistická, navazuje na předchozí tradici)
- **algebraická** (matematické metody)
- **strojová** (počítačová)

# Joshua Whatmough (1897–1964)

- britský jazykovědec (obecná a komparativní lingvistika)
- studoval klasickou filologii na Manchesteru (bakalář v roce 1916)
- potom se zúčastnil první světové války
- absolvoval postgraduální studium (v roce 1921 magistr na Manchesteru a bakalář na Cambridge)
- nejprve pracoval jako lektor (přednášející klasické filologie na Univerzitě College v severním Walesu)
- 1925–1926 zastával funkci profesora latinského jazyka na univerzitě v Káhiře
- v roce 1926 se přestěhoval do USA, kde strávil zbytek svého života
- profesor komparativní filologie na Harvardu, 1951 *Department of Linguistics* (statistické a matematické metody)

# Joshua Whatmough



- „*I'm not mad. I'm eccentric.*“
- rád užíval citáty antických myslitelů a cizojazyčné nadávky (uměl 8–22 jazyků)
- zdravě vyhlížející stařík o holi, s čerstvou chrpou v klopě
- celý život vstával před 4. hodinou ranní, jeho výuka trvala jen 30 minut
- poslal víc stížností než celý Harvard dohromady

## VIII. mezinárodní lingvistický kongres v Oslo 5.–9. 8. 1957

- předsedou kongresu byl norský lingvista A. Sommerfelt
- hojná účast ze všech zemí
- účast zástupců evropského a amerického strukturalismu (Bloomfield, Hjelmslev, Jacobson)
- delegaci ČSAV tvořili: Bohuslav Havránek, Václav Machek, Karel Horálek a Eugen Pauliny

# VIII. mezinárodní lingvistický kongres v Oslo

## 5.–9. 8. 1957

- **obecně lingvistická témata** (jazyková struktura a analýza, vztahy jazykových plánů, typologie, lexikologie, lexikografie)
- **fonetika** (instrumentální fonetika, překladové stroje)
- **srovnávací studia indoevropských jazyků** (zejména chetitština a tocharština)
- J. Whatmough referoval o existenci **Semináře matematické lingvistiky** na Harvardu (od r. 1955) – poprvé na akademické půdě a o rozšíření matematické lingvistiky na jiné univerzity v Evropě (Bonn), Spojených státech (MIT) a v Sovětském svazu (Moskva a Leningrad)