PLIN041 Vývoj počítačové lingvistiky
Kvantitativní lingvistika I

Mgr. Dana Hlaváčková, Ph.D.

Formality

- prezentace v Učebních materiálech IS
- ELF – doplňující materiály
- požadavek na ukončení – krátký referát

Pár poznámek na úvod

- přibližně **100 let** vývoje oborů, které jsou dnes součástí počítačové lingvistiky (90. léta 19. st. až 90. léta 20. st.)
- **USA**, **Evropa**, část **Asie** (Sovětský svaz)
- armádní a vládní organizace, univerzity, soukromé nadace, soukromí investoři
- polovina z této doby – **práce bez počítačů** (do 2. sv. v.)
- dvě světové **války**, studená válka, **ideologie** na obou stranách železné opony (zákaz práce a publikování, emigrace)

Pár poznámek na úvod

- dnešní počítačová lingvistika není jednotný obor
- **lingvistika** spojená s **přírodními** (matematika, informatika) a **technickými** vědami (konstrukce počítačů)
- během vývoje neustálený název
  - matematická lingvistika, jazykové inženýrství, strojová lingvistika
- dnes **Počítačová lingvistika**, **Computational Linguistics**
- **výjimečné osobnosti** s širokým záběrem vědních oborů (často nejde o lingvisty)
- **zdroje informací** – knihy, články, vzpomínky studentů, nekrology (angličtina, čeština, francouzština, ruština)
- seznam literatury v ELFu

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.)

- **posun lingvistiky k přírodním vědám**
- Historicko-srovnávací jazykověda (19. st.)
- Mladogramatická škola (konec 19. st.)
- **experimentální fonetika** (rozvoj přírodních věd, přístrojové vybavení)
- nové směry počátku 20. st., **vliv přírodních věd**
- Ferdinand de Saussure a jeho vliv v lingvistice
- strukturalismus a americký deskriptivismus
- raná korpusová lingvistika (bez počítačů)

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

- 19./20. st. – pronikání statistických metod do lingvistiky
- centrem pozornosti – **frekvence**
- zpočátku zájem na praktické úrovni – tiskaři, stenografové, Morseova abeceda
- německý stenograf **F. W. Käding** – *Slovník četnosti výskytu německého jazyka*, *Häufigkeitswörterbuch der deutschen Sprache,* 1897–98
  - cca **11 mil. slov** (právní a obchodní texty)
  - první frekvenční slovník na světě
  - frekvence slov, slabik a písmen
  - 320 výrazů pokrývá ¾ 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 rozhodovací 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 . f = k, čím je rank slova nižší, tím je jeho frekvence vyšší (Těšitelová – platí pro střední část frekvenčního slovníku)

- **2. Zipfův zákon**

a . 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.)

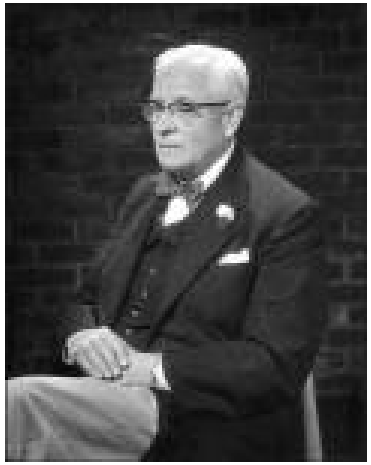
- po 2. sv. v. vznik pomezních disciplín (matematická lingvistika, sociolingvistika, psycholingvistika, etnolingvistika atd.)
- přelom 50. a 60. let, **1957 VIII. mezinárodní lingvistický kongres v Oslo**
- Joshua Whatmough **–** Mathematical Linguistics
- matematická lingvistika – využívání metod přírodních věd (statistika, algebraické metody), seminář matematické lingvistiky na Harvardu od r. 1955

- **kvantitativní** (statistická, navazuje na předchozí tradici)
- **algebraická** (matematické nekvantitativní metody)
- **strojová** (počítačová)

Joshua Whatmough (1897–1964)

- britský jazykovědec (obecná a komparativní lingvistika)
- studoval klasickou filologii na univerzitě v Manchesteru (bakalář v roce 1916)
- potom se zúčastnil první světové války
- absolvoval další studium (v roce 1921 magistr na Manchesteru a bakalář na Cambridge)
- nejprve pracoval jako lektor (přednášel klasickou filologii na univerzitě 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.“*

- pionýr matematické lingvistiky
- 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
- oblíbený učitel, i když jeho výuka byla náročná
- 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
- kongres se poprvé konal v roce 1928, koná se dodnes každých pět let
- letos **21st International Congress of Linguists** (ICL) v polské Poznani

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)
- **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)