PLIN041 Vývoj počítačové lingvistiky

Mgr. Dana Hlaváčková, Ph.D.

Matematická lingvistika I

Kvantitativní lingvistika

Frekvence a statistika

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.
- 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 (sanskrt) William D. Whitney – frekvence anglických hlásek
- německý stenograf **F. W. Käding** – *Slovník četnosti výskytu německého jazyka*, 1897–98, přes 10 mil. slov (*Häufigkeitswörterbuch der deutschen Sprache*) první frekvenční slovník, frekvence slov, slabik a písmen, 320 výrazů pokrývá ¾ textu

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

- ruský matematik Andrej Andrejevič Markov – *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 . f = k, čím je rank slova nižší, tím je jeho frekvence vyšší (Těšitelová – platí pro střední část frekvenčního slovníku)

- **2. Zipfův zákon**

a . b = k, čím je frekvence nižší, tím více slov tuto frekvenci má

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

Frekvenční slovníky

- 1. pol. 20. st. – 70. léta
- potřeby stenografie a didaktiky (výuka cizích jazyků)
- lexikální statistika (jazyk, text, dílo, autor…)
- frekvenční seznamy (typy třídění)
- frekvenční slovníky (informace o slovní zásobě)
  - rozsah (500 tis. – 11 mil. slov)
  - výběr zpracovaných textů
  - technika zpracování (ruční, strojové)
- různé počty slovníků v jednotlivých jazycích

Frekvenční slovníky – předzvěst korpusů

- rozsáhlý jazykový materiál
- frekvenční studie
- frekvence z hlediska morfologie, syntaxe, sémantiky
- stylová rozrůzněnost (vyváženost)
- mluvený jazyk
- definice „slova“
- otázka homonymie

Frekvenční slovníky - němčina

- návaznost na F. W. Kädinga (data pro jiné účely)
- Bayard Quincy Morgan – *German Frequency Word Book*, 1928 (New York), pedagogické účely, 2400 nejčastějších slov
- Helmut Meier – *Deutsche Sprachstatistik I/II*, 1964
- Hans Heinrich Wängler (fonetik) – *Rangwörterbuch hochdeutscher Umgangssprache*, FS hovorové horní němčiny (denní tisk + magnetofonové nahrávky a jejich transkripce), 1963
- Inger Rosengren, frekvence slovní zásoby z novin *Die Welt* (6 mil.) a *Süddeutsche Zeitung* (6 mil.), 5 tematických kategorií z let 1966–1967, 1972

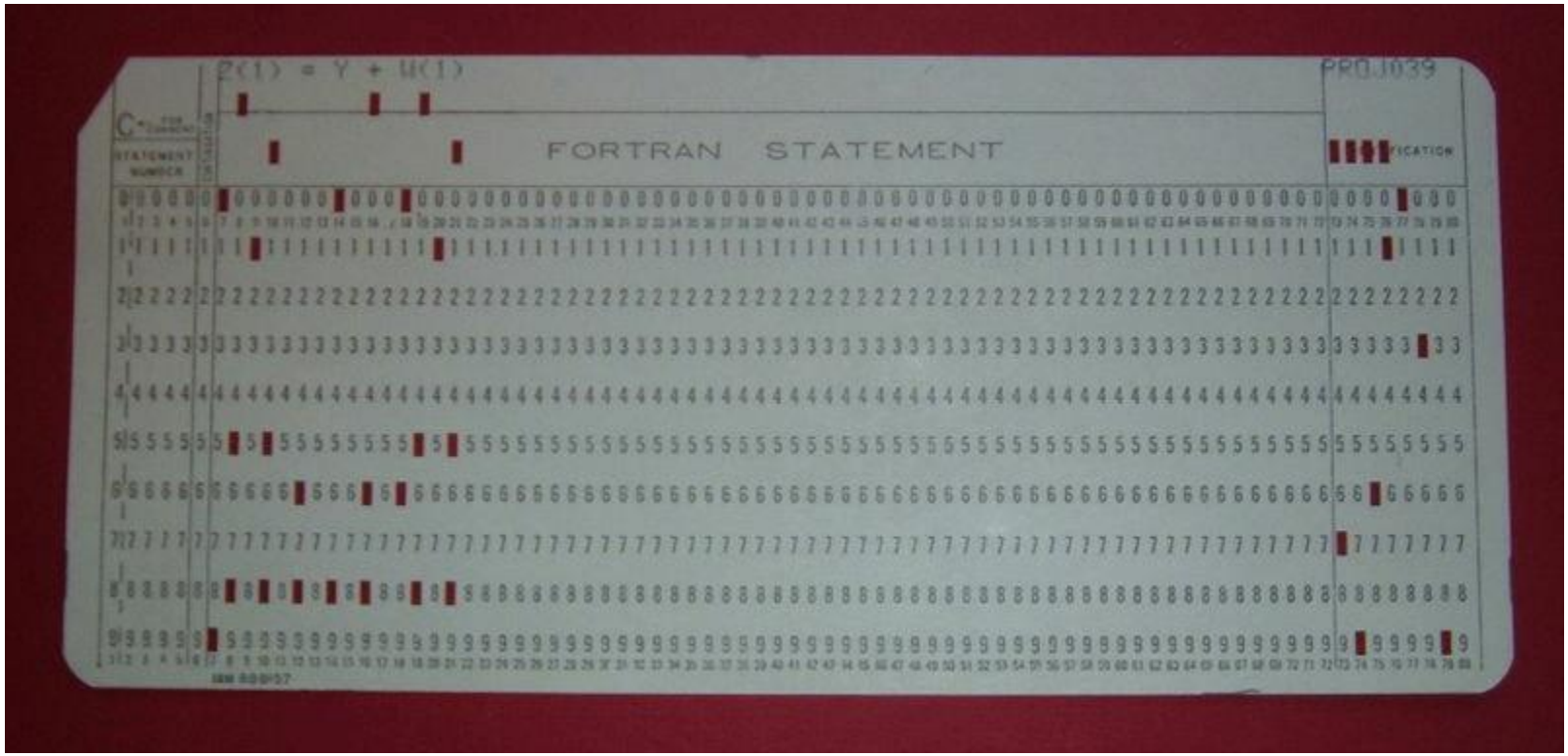
Frekvenční slovníky – angličtina

- L. P. Ayres – *A Measuring Scale for Ability in Spelling*, obchodní a soukromé dopisy, 1915, 368 000 slov
- **Edward Lee Thorndike** (psycholog, *Animal Intelligence*, 1911) – *The Teacher's Word Book*, 3 díly, 1921, 1932, 1944
- Michael West – *A General Service List of English Words with Semantic Frequencies*, (GSL) 1973
- **Henry Kučera** (1925–2010), filozofie a lingvistika na UK, po 1948 emigrace, Brown University, W. Nelson Francis – ***Brown Corpus of Standard American English***, 1964, americká angličtina, 1 mil. slov

Computational Analysis of Present-Day American English, 1967

- použití počítačů IBM, děrné štítky a magnetické pásky

Děrný štítek



Henry Kučera, čestný doktorát na MU, 1990

Frekvenční slovníky – románské jazyky

- Alphonse Juilland (1923–2000)
- španělština – *Frequency Dictionary of Spanish Words*, 1964
- rumunština – *Frequency Dictionary of Rumanian Words*, 1966
- francouzština – *Frequency Dictionary of French Words*, 1970
- italština – *Frequency Dictionary of Italian Words*, 1973
- 5 různých žánrů, 500 tis. slov, 1920–1940, strojové zpracování

Frekvenční slovníky – ruština

- Harry Hirsch Josselson – *The Russian Word Count and Frequency Analysis of Grammatical Categories of Standard Literary Russian*, Detroit 1953, 1 mil. slov z umělecké literatury let 1830–1950
- E. A. Šteinfeldt – *Častotnyj slovar' sovremennogo russkogo literaturnogo jazyka*, 1963, dětská literatura, pro výuku ruštiny na estonských školách (2500 nejčastějších slov)
- Lidija Nikolajevna Zasorina a kol. – *Častotnyj slovar' russkogo jazyka: okolo 40 000 slov*, 1977, ruština 20. st. (Lenin, Gorkij, Šolochov, XII. a XIII. sjezd KSSS, novinové články 1968)
- *Slovar' jazyka Puškina I–III*, 1956–1961, Viktor Vladimirovič Vinogradov (ed.), jazyk a styl klasických ruských autorů, spisovný ruský jazyk

Frekvenční slovníky – čeština, slovenština

- Jozef Mistrík (1921–2000) – Frekvencia slov v slovenčine, 1969
- jazykovědec, literární vědec, pedagog, soudní grafolog
- stylistika – funkční styly, teorie komunikace
- od r. 1965 na Filozofické fakultě UK Bratislava (oddělení matematické lingvistiky, Katedra slovenského jazyka)
- stenografie (1954–1960 ředitel Štátneho stenografického ústavu v Bratislavě, např. těsnopis pro nevidomé)
- Retrográdny slovník slovenčiny, 1976
- Frekvencia tvarov a konštrukcií v slovenčine, 1985

Frekvencia slov v slovenčine (FSS)

- možnost srovnání české a slovenské frekvence slov, FSČ a FSS mají však rozdílné parametry
- velikost 1 mil. slov
- základní lexikální jednotkou je grafické slovo (složené slovesné tvary)
- výběr 60 textů – 5 stylových skupin (dialogy, umělecká próza, poezie, žurnalistika, naučná literatura), nevyváženost (více textů od jednoho autora), nestejná délka textů
- frekvenční seznam 9 568 slov do frekvence 3
- (dnes PhDr. Mária Šimková vedoucí oddělení SNK na JÚĽŠ)

FSČ

- Jaroslav Jelínek, Josef V. Bečka, Marie Těšitelová – Frekvence slov, slovních druhů a tvarů v českém jazyce, 1961
- *viz samostatnou prezentaci*

- František Čermák, Michal Křen (eds.) – Frekvenční slovník češtiny, 2004 – založen na korpusu FSC2000, 95 mil. slovních tvarů

Kvantitativní vztahy v jazyce za základě FS

- 3 pásma frekvenčního seznamu (nejvyšší, střední, nejnižší)
- v 1. pásmu leží 10 nejfrekventovanějších slov – velmi krátká slova, pokrývají cca 20 % textu (1. slovo 5 % „a“)
- formálních slov je málo s vysokou frekvencí *(koncentrace slovníku)*, plnovýznamových slov je hodně s nízkou frekvencí *(bohatství slovníku), např. v češtině 20:80, ve francouzštině 50:50*
- koeficient *disperze (rozptýlení = rozdělení frekvence slov v různých textech), 0 rovnoměrné–1 nerovnoměrné*