PLIN041 Vývoj počítačové lingvistiky
Kvantitativní lingvistika II

Mgr. Dana Hlaváčková, Ph.D.

Frekvence a statistika
Frekvenční slovníky

2. pol. 19. st. – 60. léta 20. st.

Frekvenční slovníky

- 1. pol. 20. st. – 70. léta
- potřeby stenografie a didaktiky (výuka cizích jazyků), postupně zájem lingvistiky
- **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

- řeší se stejné problémy jako později u korpusů
- **rozsáhlý** jazykový materiál
- frekvence z hlediska **morfologie**, **syntaxe**, **sémantiky**
- otázka **homonymie**
- definice „**slova**“ (slovníkové heslo, lemma)
- stylová rozrůzněnost (**vyváženost**)
- **mluvený jazyk**

Frekvenční slovníky – němčina

- návaznost na F. W. Kädinga (data pro jiné účely)
- Bayard Quincy Morgan (prof. němčiny a překladatel na Stanford University) – *German Frequency Word Book*, 1928 (New York), pedagogické účely, 2400 nejčastějších slov
- Helmut Meier (germanista) – *Deutsche Sprachstatistik I/II*, 1964, nejfrekventovanější slova z Kädinga, cca 10 tis. slov
- 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

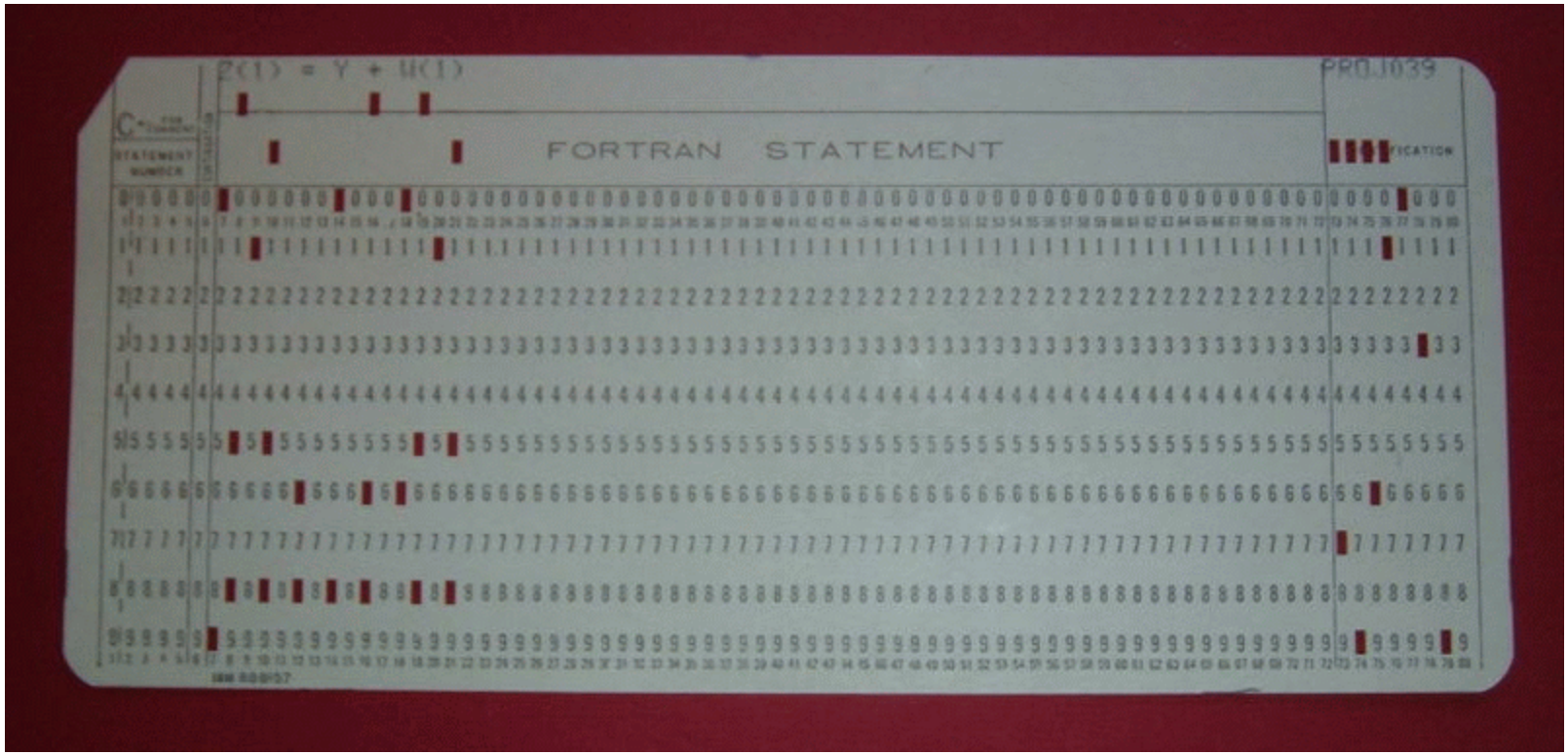
Frekvenční slovníky – angličtina

- L. P. Ayres – *A Measuring Scale for Ability in Spelling*, obchodní a soukromé dopisy, 1915, 368 000 slov
- **Edward Lee Thorndike** (psycholog, proces učení, *Animal Intelligence*, 1911) – *The Teacher's Word Book*, 3 díly, 1921, 1932, 1944
- Michael West – *A General Service List of English Words*, (GSL) 1953, 2 000 slov, obnovován do současnosti – NGSL
- **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), pův. Rumun, studoval v Paříži romanistiku, působil v USA
- 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 – slovenština

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

- 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**
- výběr **60 děl** – **5 stylových skupin** (jevištní dialogy, umělecká próza, poezie, žurnalistika, naučná literatura), nevyváženost (více textů od jednoho autora), nestejná délka textů
- literatura z let **1922–1966**
- frekvenční seznam 9 568 slov do frekvence 3
- (později dr. Mária Šimková z oddělení SNK v JÚĽŠ)

FSČ

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

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

FSČ

- peripetie vzniku díla po druhé světové válce
- práce byla započata v r. **1940** pod vedením Vladimíra Šmilauera v *Kruhu přátel českého jazyka* a dokončena **1953**
- po roce 1948 musely být některé texty (např. projevy Edvarda Beneše) vypuštěny a jiné (např. projevy Antonína Zápotockého) nově zařazeny
- hrozilo, že se slovník dostane do rozporu se Stalinovým učením o marxismu v jazykovědě
- prosadil ho až **Jaromír Bělič** v roce 1961; napsal předmluvu
- obsahuje kapitolu o nejvýznačnějších pracích o frekvenci slov (stejně jako FSS)

FSČ

- velikost slovníku – **1 623 527 slov**
- je založen na výpiscích ze **75 děl**
- materiál je rozdělen do **8 funkčních oblastí** užívání jazyka
  - čtyři umělecké (beletrie, poezie, literatura pro mládež, dramata)
  - čtyři odborné (odborná lit., žurnalistika, vědecká lit., mluvené projevy uveřejněné tiskem)
- měly být zařazeny soukromé dopisy a mluvené projevy (obtíže se zaznamenáváním)
- díla vydaná po roce **1930**
- frekvence slov
- frekvence typů skloňování substantiv
- frekvence pádů, rodů a čísel při skloňování
- frekvence slovesných tvarů

FSČ

- nestejné zastoupení: beletrie 30 %, mluvené projevy 6 %
- problém s **mnohoznačností** slov, není zachycena (např. *hlava* = lidská, rybí, *ztrácet hlavu*, *hlava státu*, *hlava šroubu* apod.)
- těžkosti s pojmem „**slovo**“ (např. *byl bych býval přišel*)
  - jednotlivá slova (předložky, členy sousloví např. *střední (škola)*)
  - víceslovná hesla (slovesné tvary např. *přišel jsem*, sloveso se zvratným zájmenem např. *mýti se*, spřežky psané zvlášť i dohromady)
- každé slovo se zapisovalo na kartotéční lístek formátu A5, zachycovala se celková frekvence a frekvence v jednotlivých stylových oblastech a počet pramenů, ve kterých se vyskytlo

FSČ

pořadí slova
frekvence slova:
- **absolutní frekvence**
- **počet skupin**, v nichž se slovo vyskytlo
- **počet pramenů**, v nichž je heslo doloženo

	a (spoj.)	67122-8-75	
	býti	43148-8-75	
	ten	37280-8-75	
	v(e)	33679-8-75	
	on	32496-8-75	
	na	27753-8-75	
	že	18092-8-75	
	s(e)	14951-8-75	
	z(e)	13408-8-75	
10	který	11692-8-75	
	míti (se)	11426-8-75	
	já	11060-8-64	
	k, ke, ku	11038-8-75	
	do	10831-8-75	
	i	10559-8-75	
	ale	9406-8-73	
	svůj	9121-8-74	
	jako	8896-8-75	
	o (předl.)	8641-8-75	
20	tak	7930-8-75	
	co	7907-8-75	
	za	7712-8-75	
	se (zájm.)	7174-8-75	
	tento	6841-8-71	
	moci	6763-8-74	
	jen	6477-8-75	
	po	6338-8-75	
	aby	6282-8-75	
	jak	6123-8-74	
30	jenž	5956-8-73	

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 údajně 50:50 (Pierre Guiraud)*
- koeficient *disperze (rozptýlení = rozdělení frekvence slov v různých textech), 0 rovnoměrné–1 nerovnoměrné*