PLIN041 Vývoj počítačové lingvistiky

Strojová lingvistika

Strojový překlad

Mgr. Dana Hlaváčková, Ph.D.

40.–70. léta 20. století

Strojová lingvistika

- praktická aplikace *kvantitativní* a *algebraické* lingvistiky
- nárůst množství informací
- snaha o automatické zpracování informací a textů
- uložení a opětovné nalezení informace – bibliografické údaje, kartotéky, knihovnické automatizované systémy – klasifikace, indexování
- informace o obsahu – získávání informací (Information Retrieval), terminologie, klíčová slova, sestavování konkordancí
- práce s informacemi ve dvojkové soustavě – je možné zakódovat jazykové kategorie, které mají binární povahu
- algoritmický popis jazyka

Strojový překlad (od 50. let 20. st.)

- první písemná zmínka 1947 – dopis W. Weavera N. Wienerovi:

„… I have wondered if it were unthinkable to design a computer which could translate.“

„This is really written in English, but it has been coded in some strange symbols“ (při pohledu na článek v ruštině, kryptografie)

1949 – Weaver – memorandum *Translation*

1952 – *The first machine translation conference* (MIT, Bar-Hillel)

1954 – *Mechanical Translation* (journal)

1955 – *Machine Translation of Languages: Fourteen Esays*

- strojový překlad je záležitost kryptografie
- snaha o formalizaci jazyka pro strojové zpracování, vytvoření *převodního jazyka* = formalizace významové struktury vět (nevyřešeno)

Warren Weaver 1894–1978

- americký vědec, matematik, „pionýr“ strojového překladu, organizátor vědy
- vystudoval univerzitu ve Wisconsinu, profesor matematiky

v Kalifornii

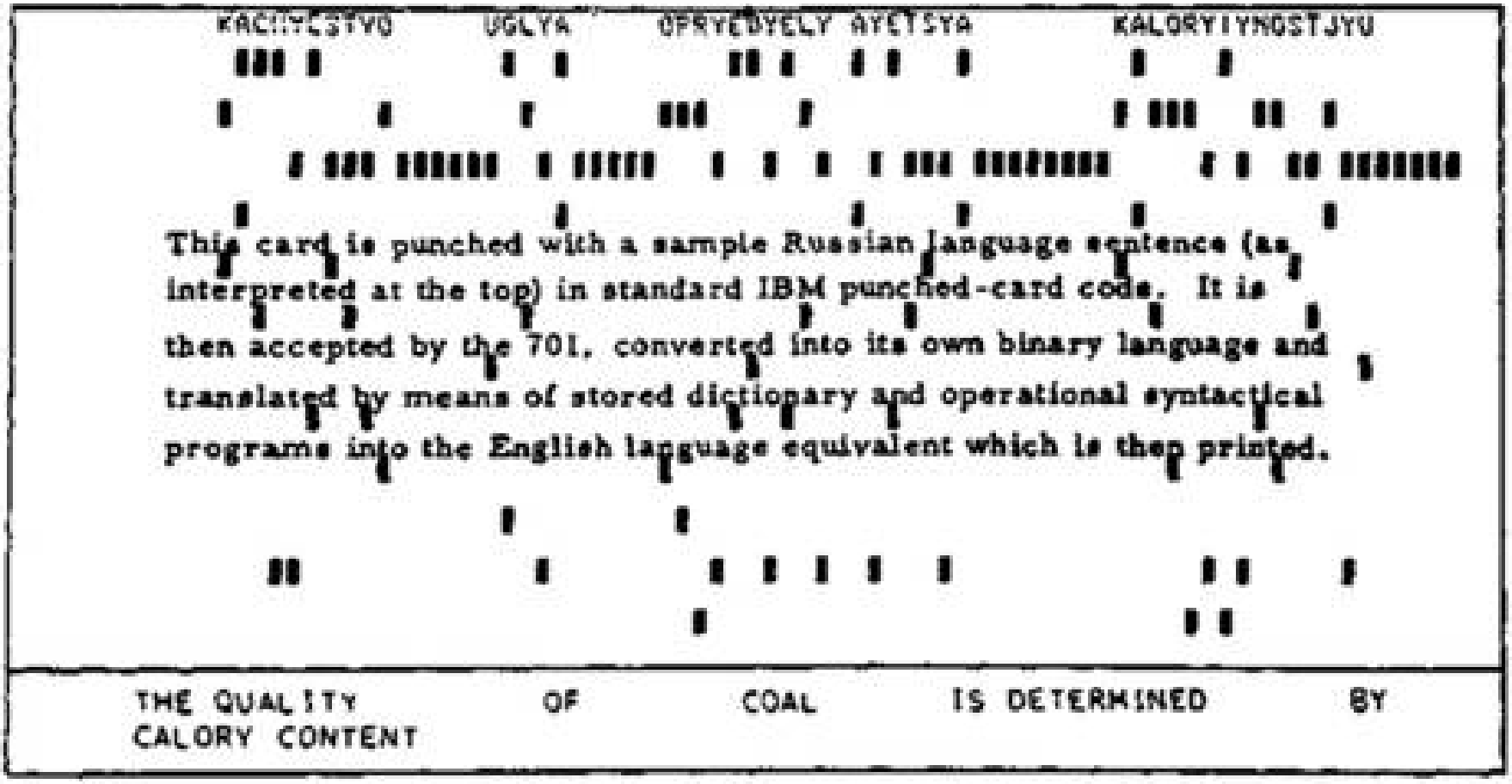
- sloužil v 1. sv. v. u letecké služby
- 1932–1955 byl ředitelem oddělení přírodních věd Rockefellerovy nadace, zal. 1913 (Division of Natural Sciences Rockefeller Foundation), přidělování grantů v přírodních vědách
- za 2. sv. v. – řídil operace matematiků, zabýval se počítacími stroji a kryptografií
- iniciátor strojového překladu
- nadační činnost – podpora řady vědeckých projektů
- sběratel *Alenky v říši divů* (v r. 1964 – 160 verzí ve 42 jazycích)

Strojový překlad (pol. 50. – poč. 60. let)
1) nadšení a velké investice

- převod textu ze vstupního jazyka (analýza) do výstupního jazyka (syntéza)
- motivace – studená válka, narůstající počet textů v různých jazycích
- vysoká chybovost strojových překladů
  - malý rozsah paměti počítačů
  - překlad slovo za slovo (bez znalosti syntaxe)
  - vliv mimojazykových skutečností
- hledání *univerzálního převodního jazyka* – symbolická konstrukce popisující významovou strukturu věty
  - problém s formalizací významů
- pouze jednosměrné překlady z jednoho jazyka do druhého (často mezi angličtinou a ruštinou)

Georgetown-IBM experiment

- 7. 1. 1954, v centrále IBM v New Yorku, počítač IBM 701
- Cuthbert Hurd – ředitel Applied Sciences Division v IBM
- Peter Sheridan – programátor IBM
- Léon Dostert – za 2. sv. v. tlumočník generála Eisenhowera, po válce – návrh simultánního překladu (sluchátka) během Norimberského procesu
- Paul Garvin – lingvista (pův. Čech)
- Institute of Languages and Linguistics Georgetown University
  - 49 vět z ruštiny do angličtiny
  - organická chemie, obecné věty
  - 137 lexikálních jednotek (kmeny a koncovky)
  - 6 gramatických pravidel



Specimen punched card and below a strip with translation, printed within a few seconds

Strojový překlad (1. fáze)

- Erwin Reifler (1903–1965) University of Washington, Far Eastern Department
- původem Rakušan, profesor čínštiny, „pionýr“ strojového překladu, 1932–1947 Čína, Hong Kong, Japonsko (studium čínských znaků)
- 1948 článek *The Chinese Language in the Light of Comparative Semantics* – četl Weaver a poslal Reiflerovi své memorandum
- nadšení pro MT
- 1951 série článků *Studies in Mechanical Translation*
- The Machine Translation Project, sponzorován letectvem USA, angličtina – ruština, angličtina – čínština

Strojový překlad (pol. 60. – poč. 70. let)
2) rozčarování

- ALPAC (Automatic Language Processing Advisory Committee), výbor ustanoven 1964 (7 vědců) – zhodnocení pokroku ve strojovém překladu
- zpráva ALPAC z r. 1966, velká skepse k MT, doporučení používat počítače v překladu jen jako praktickou pomůcku a budovat elektronické slovníky
- velké investice a zklamání
- zrušení investic i výzkumu
- soustředění pozornosti na budování slovníků a výzkum jazyka (gramatika, sémantika)
- později se vyplatil bližší výzkum jazyků

Strojový překlad (70. a 80. léta)
3) umírněný optimismus a částečné úspěchy

- do popředí se dostává Kanada, Francie, SSSR, Japonsko
- Kanada, Montreal 1975 – systém TAUM-METEO (autor John Chandioux)
  - první automatický systém
  - překlad anglických meteorologických zpráv do francouzštiny
  - jednoduchá slovní zásoba i gramatika, 10–20 % svěřeno překladateli, jazyk Q, cca 1 000 slov/min
- Francie, Grenoble (Bernard Vauquois) – skupina GETA, překlady z ruštiny do francouzštiny, jazyk ALGOL
- Japonsko – univerzity Tokio, Osaka, firma Hitachi
  - **Makoto Nagao** – průkopník NLP a MT v Japonsku
  - úspěšný překlad mezi angličtinou a japonštinou
  - example-based MT, rozdělení na fráze

Yehoshua Bar-Hillel

- 1915 Vídeň – 1975 Jeruzalém
- izraelský filozof, logik, matematik, lingvista
- předmět zájmu – algebraická lingvistika, strojový překlad, získávání informací, logické aspekty přirozených jazyků
- 1933 emigroval do Palestiny, žil v kibucu, usadil se v Jeruzalémě
- za 2. sv. v. bojoval v Židovské brigádě britské armády, potom v Izraelské válce o nezávislost (přišel o oko)
- Hebrew University v Jeruzalémě střídá s Research Laboratory of Electronic na MIT (strojový překlad)

Yehoshua Bar-Hillel

- snaha o sblížení logiky a lingvistiky (sémantiky)
- spolupráce s R. Carnapem, formální popis sémantiky, článek *Semantic Information*, 1953
- kategoriální gramatika (vliv Chomského)
- od poč. 50. let ovlivněn kybernetikou (N. Wiener), práce na MIT v oblasti strojového překladu
- 1952 vede první konferenci o strojovém překladu
- velké nadšení, ale zdůrazňuje podíl lidské práce při překladech
- konec 50. let – skepse a silná kritika využití statistických metod a hledání mezijazyka, ani rozsáhlá data nevyřeší všechny víceznačnosti

Yehoshua Bar-Hillel

- 1960 zpráva sponzorům a vládě USA o neúspěšnosti MT (bývala citována jako důkaz nemožnosti MT)
- *Language and Information*, 1964
- *Aspects of Language*, 1970
- (vnučka Gili Bar-Hillel přeložila Harryho Pottera do hebrejštiny)

Andrew Donald Booth

- 1918–2009
- matematický fyzik, elektroinženýr,
- za 2. sv. v. – krystalografie (výbušnin)
  - mechanické a elektromechanické kalkulátory (pro triviální aritmetické výpočty)
- po 2. sv. v. – 4 skupiny v Británii – budování počítačů
- Booth – Birkbeck College University of London (musí být dost levný, levné komponenty)
- 1946 Rockefellerova nadace – cesta po USA za návrháři prvních počítačů (John von Neumann, Princeton) – návrh počítače s von Neumanovou architekturou
- návrh paměti – magnetický buben
- 1947 setkání s Weaverem – financování výzkumu krystalografie (electronic computer), strojový překlad
- počítač Automatic Relay Computer (ARC)