Matematická lingvistika II

Algebraická lingvistika

Strojová lingvistika

Strojový překlad

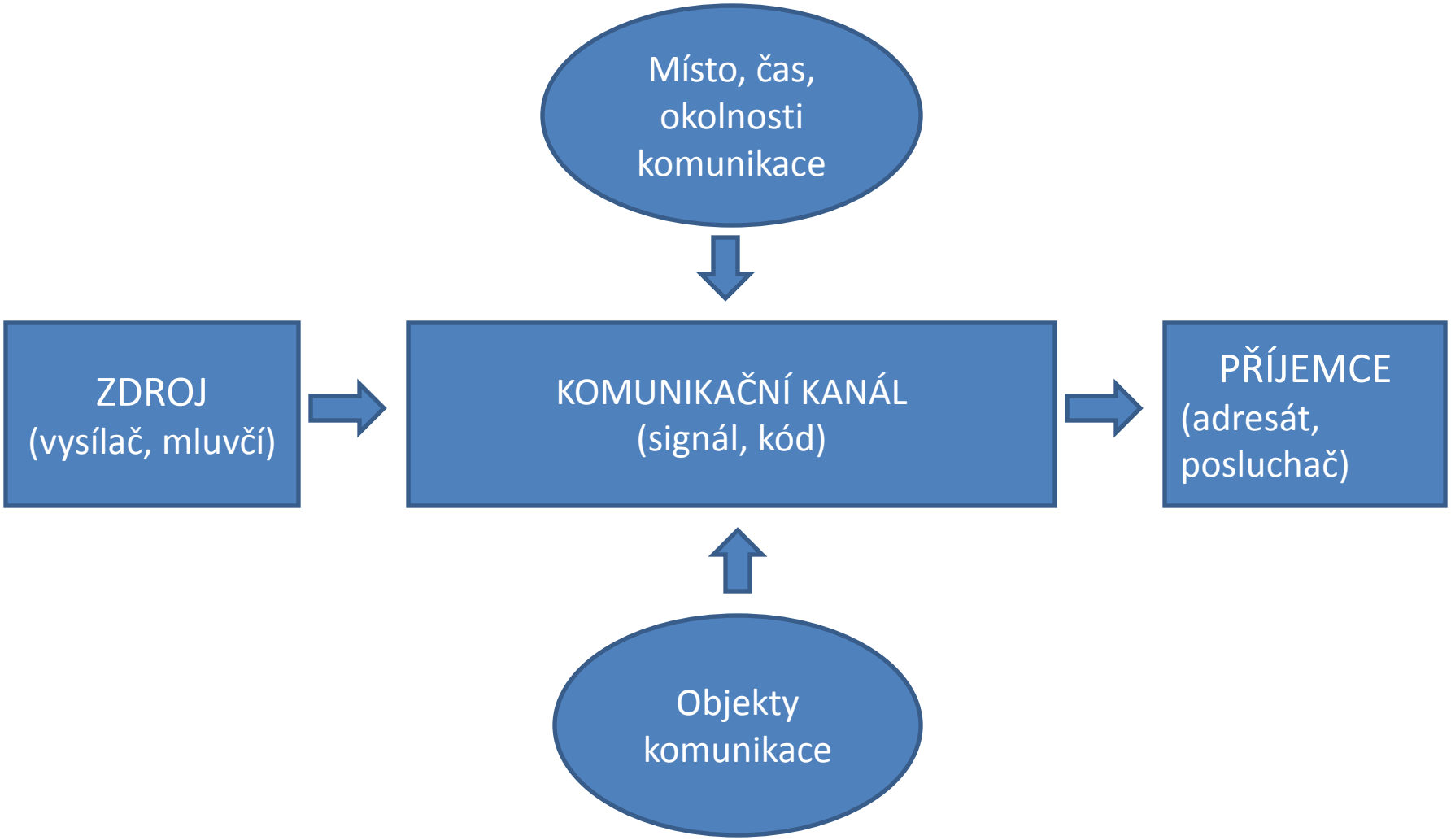
(teorie komunikace, teorie informace)

40.–70. léta 20. století

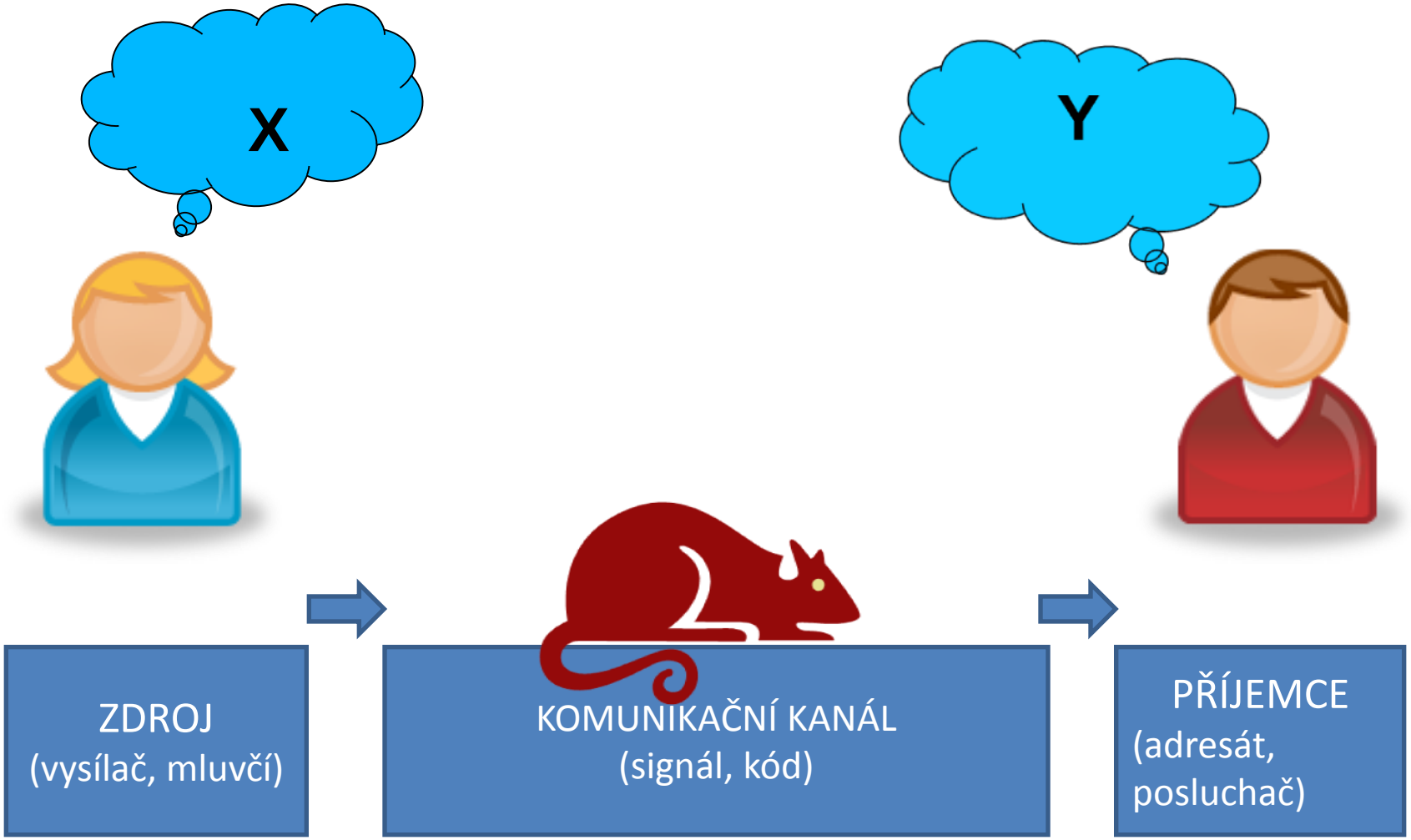
Teorie komunikace a informace

- 40./50. léta, přenos informace, vznik kybernetiky
- Claude Elwood Shannon (angl. matematik), Warren Weawer (am. matematik, fyzik) – The Mathematical Theory of Communication, 1949
- Charles Francis Hockett (am. strukturalista) – Review of Shannon & Weawer, Language, 1953
- komunikace – mluvčí + kód (zakódování informace) – kanál – příjemce + kód (dekódování)

Model jazykové komunikace



Komunikační šum



Teorie informace

- mluvčí – myšlenky → zvukový signál
- příjemce – na základě dosud dekódované výpovědi odhaduje další část (pravděpodobnost, Markovův proces)
- množství informace se dá měřit – entropie – *průměrné množství informace připadající na jeden komunikační znak*
- entropie je tím větší, čím je znak méně předvídatelný – předvídatelnost (predictability) – míra pravděpodobnosti, s jakou příjemce odhadne další část výpovědi
- nulová entropie = redundance, spolehlivost přenosu x šum
- vztah entropie a frekvence (nižší frekvence vyšší entropie)
- míra informace je individuální – zkušenost, vzdělání, věk
- jednotka informace – bit (binary digit), binární opozice 0/1

Teorie informace

- teorie informace – kybernetika – strojová lingvistika – strojový překlad
- sborník **Teorie informace a jazykověda**, 1964 (překlady zásadních článků z této oblasti)
- Roland Lvovič Dobrušin – Matematické metody v lingvistice, 1961 – zdokonalení strojového překladu
- Warren Plath – Matematická lingvistika, 1961 – přehled dosavadního vývoje
- V. V. Ivanov, S. K. Šaumjan (přední sovětský strukturalista) – Lingvistické problémy kybernetiky a strukturní lingvistika, 1961, *Kibernetiku na službu kommunizmu*
- C. E. Shannon – Predikace a entropie tištěné angličtiny, 1951

Teorie informace

- Benoît Mandelbrot – Komunikace a formální struktura textů, 1954, vliv fyzikálních a fyziologických podmínek na komunikaci, francouzský matematik, zakladatel fraktální geometrie
- Vitold Belevitch – Teorie informace a lingvistická statistika, 1956, vztah délky slova a množství informace; belgický matematik ruského původu
- Yehoshua Bar-Hillel (izraelský filozof, matematik, lingvista), Rudolf Carnap (německý filozof, matematik, logik; novopozitivismus, teorie vědy) – Sémantická informace, 1953, teorie sémantické informace, důležitý je význam informace; strojový překlad

Teorie informace

- Paul L. Garvin – Stupně začlenění počítačů do lingvistického výzkumu, 1962, strojový překlad; americký lingvista (původem Čech), sociolingvista, antropologická lingvistika, 1990 čestný doktorát MU
- S. M. Lamb – Číslicový počítač jako pomocník v lingvistice, 1961, IBM 650, IBM 704 univerzity v USA (*MIT, Michigan, Washington, Berkeley, Los Angeles, Harvard, Pennsylvania, Severní Karolina*), nájem 45 tis. dolarů měs., 1 min 6 dolarů

Algebraická lingvistika
(2. pol. 50. let 20. st.)

- rozvoj matematiky a formální logiky
- hledání jazyka vědy (přirozený jazyk – nevhodný, 2. pol. 19. st.) – matematická, formální logika (konjunkce, disjunkce, implikace, ekvivalence)
- využití nekvantitativních matematických metod v lingvistice, formální lingvistika (formální popis gramatik a jazyků)
- algebraická lingvistika – Y. Bar-Hillel
- význam pro strojový překlad a strojovou lingvistiku

Matematické modely v lingvistice

- generativní transformační mluvnice (Chomsky) – generování gramaticky správných vět z výchozího symbolu na základě souboru pravidel
- kategoriální (rekognoskativní) gramatika (Bar-Hillel) – věta = řetězec symbolů, zjišťuje se struktura věty a gramatická správnost
- aplikačně-generativní model (Šaumjan) – jazykové jednotky (symboly) a vztahy mezi nimi se odvozují metodami matematické logiky (spojení strukturalismu a generativní gramatiky; genotypický a fenotypický jazyk)

Matematické modely v lingvistice

- teorie analytických modelů (analytická metoda)
- pro slovanské jazyky – rozvinutá morfologie a volný slovosled
- teorie množin – výchozí množina = gramaticky správné věty; podmnožiny = soubory základních jednotek (lexikologie, morfologie, syntax; syntagma i paradigma)
- SSSR – Olga Sergejevna Kulagina, R. L. Dobrušin, I. I. Revzin, I. A. Meľčuk
- Rumunsko – Solomon Marcus
- ČSSR – Ladislav Nebeský (teorie grafů, binární básně)
- *Analytický směr v algebraické lingvistice* (SaS 1967, s. 161–7)

Matematické modely v lingvistice

- závislostní gramatika a teorie grafů (nelineární zobrazení věty, graf)
- američtí lingvisté – David G. Hays, K. E. Harper; strojový překlad (*Použití strojů při konstruování gramatiky*, 1959)
- sovětští lingvisté – D. S. Cejtin, L. N. Zasorina
  (*O vyčlenění konfigurací v ruské větě*, 1961)
- *závislostní syntax* (druhotně morfologie, jiné jazykové roviny ne)
  - závislostní pravidla
  - závislostní strom – uzel a hrany
- ČSSR – funkční generativní popis (Panevová, Sgall)

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 Essays*

- snaha o formalizaci jazyka pro strojové zpracování, vytvoření „převodního jazyka“ = formalizace významové struktury vět (nevyřešeno)

Strojový překlad (od 50. let 20. st.)

- 1) nadšení a velké investice, motivace – studená válka (pol. 50. – poč. 60. let)
- chybovost – malý rozsah paměti, slovo za slovo, mimojazykové skutečnosti
- 1954 v centrále IBM Georgetown-IBM experiment (49 vět z RJ do AJ, 250 slov, 6 gramatických pravidel, IBM 701)
- University of Washington, od 1956 (Erwin Reifler – The Machine Translation Project)
- 2) rozčarování (pol. 60. – poč. 70. let)
- zpráva ALPAC (Automatic Language Processing Advisory Committee) 1966, skepse k MT (*viz P. Novák – Jazyk a stroje, SaS 1967, č. 3*)
- zrušení investic i výzkumu, soustředění pozornosti na budování slovníků a výzkum jazyka (gramatika, sémantika)

Strojový překlad (od 50. let 20. st.)

- 3) umírněný optimismus a částečné úspěchy (70. a 80. léta)
- Kanada, Montreal 1975 – systém TAUM-METEO (John Chandioux) – překlad anglických meteorologických zpráv do francouzštiny (jednoduchá slovní zásoba i gramatika, 10–20 % svěřeno překladateli, jazyk Q)
- Francie, Grenoble – RJ – FrJ, jazyk ALGOL
- SSSR
- Moskva – V. J. Rozencvejg (AJ), J. D. Apresjan (RJ), O. S. Kulagina (RJ – FrJ)
- Leningrad (Petrohrad) – G. S. Cejtin

Strojový překlad (od 50. let 20. st.)

- ČSSR – od r. 1957, Oddělení teorie strojového překladu FF UK (P. Sgall, P. Novák, B. Palek)
- r. 1960 překlad AJ – ČJ na čs. počítači SAPO (několik desítek slov, odborné texty, ustálená slovní zásoba),
- lingvisté z UK Petr Sgall, Eva Hajičová, Jarmila Panevová, Petr Piťha, Zdeněk Kirschner, Výzkumný ústav matematických strojů
  - projekt APAČ, 70. léta, ČJ – AJ
  - projekt RUSLAN, 80. léta, ČJ – RJ

Yehoshua Bar-Hillel

- 1915 Vídeň – 1975 Jeruzalém
- izraelský filozof, logik, matematik, lingvista
- předmět zájmu – algebraická lingvistika, strojový překlad, získávání informací, logické aspekty přirozených jazyků
- snaha o sblížení logiky a lingvistiky (sémantiky)
- 1933 emigroval do Palestiny, žil v kibucu, usadil se v Jeruzalémě
- za 2. sv. v. bojoval v Židovské brigádě britské armády, poté v Izraelské válce o nezávislost (přišel o oko)
- Hebrew University v Jeruzalémě střídá s Research Laboratory of Electronic na MIT (strojový překlad)
- spolupráce s R. Carnapem, formální popis sémantiky, článek *Semantic Information*, 1953
- kategoriální gramatika (vliv Chomského)

Yehoshua Bar-Hillel

- 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
- 1960 zpráva sponzorům o neúspěšnosti MT (bývala citována jako důkaz nemožnosti MT)
- kritika formálních gramatik (včetně své vlastní)
- *Language and Information*, 1964
- *Aspects of Language*, 1970
- (vnučka Gili Bar-Hillel přeložila Harryho Pottera do hebrejštiny)

Norbert Wiener

- 1894–1964, americký matematik, zakladatel kybernetiky
- *Cybernetics or the Control and Communication in the Animal and the Machine*, 1948 (Kybernetika aneb Řízení a sdělování u organismů a strojů)
- v 11 letech začal studovat na vysoké škole matematiku, v 15 letech bc. titul, Harvard – zoologie , pak filozofie, ale disertace souvisela s matematickou logikou, Ph.D. v 18 letech
- učil filozofii na Harvardu, matematiku na MIT, pracoval v oblasti balistiky
- u studentů znám chabým způsobem přednášení, vtipy a roztržitostí
- teorie pravděpodobnosti a náhodné procesy (Wienerův bílý šum)
- dodnes je udělována Wienerova cena za aplikovanou matematiku

Norbert Wiener

- za 2. sv. v. – řízené střely na velkou vzdálenost
  - automat – zasáhnout cíl, odpovědět na otázku
- na konci války – 1. radarem řízená střela, pak se věnoval 2. typu automatů
- éra kybernetických strojů, počítače, analogie s lidským mozkem (zpětná vazba na podněty z okolí u živých organismů i strojů, feedback)
- kybernetika zkoumá stroje i živé organismy
- vynutila si teorii informace
- informatika, umělá inteligence, neuronové sítě
- na východě kybernetika nejdříve buržoazní pavěda – přijata na konci 50. let

Andrew Donald Booth

- 1918–2009
- britský 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
- počítač Automatic Relay Computer (ARC)
- návrh paměti – magnetický buben
- 1947 setkání s Weaverem – financování výzkumu krystalografie (electronic computer), strojový překlad

První počítače

- Charles Babbage – angl. matematik, filozof a vynálezce – mechanický programovatelný stroj (děrné štítky pro řízení tkalcovských strojů)
- 0. generace – elektromagnetické relé (40. léta),
  - K. Zuse – Zuse Z4, H. Aiken – Mark 1 (atomová bomba)
- 1. generace – elektronky (1945–1951), pouze 1 operace
- ENIAC (1946–1955) – pro americkou armádu, turingovsky úplný, 18 tis. elektronek, poruchovost, 150 m^2, 40 t, chlazen 2 leteckými motory
- MANIAC – J. von Neumann, 1945 (vodíková bomba)
- EDVAC, Univac – Remmington
- SAPO (Ant. Svoboda) Ústřední ústav matematický, 1960

První počítače

- 2. generace – tranzistory (1952–1965),
  - dávkové zpracování, několik skříní (COBOL, FORTRAN)
- EPOS 2, 1962, A. Svoboda, 3 kusy (EPOS 1 elektronkový)
- 3. generace – integrované obvody (1965–1980)
  - 1 skříň, paralelní zpracování, IBM 360
- 4. generace – mikroprocesory (od 1981)
  - PC, zmenšování rozměrů, zvyšování operační rychlosti

Strojová lingvistika, od 60. let 20. st.

- počítače jsou pro lingvistiku potřebné a užitečné
- lingvista nemusí rozumět počítači, ale musí velmi dobře rozumět svému jazyku
- 1) lingvistický rozbor a popis jazykového jevu
- 2) formální popis jazykového jevu tak, aby mu počítač rozuměl
- s využitím počítačů se začínají rozvíjet jednotlivé oblasti počítačové lingvistiky
- lingvisté si kladou vysoké cíle – nedostatečné počítačové vybavení

Strojová lingvistika

- automatické zpracování informací a textu, *text processing*
- **uložení a opětovné nalezení informace**
  - velké množství odborné literatury, odborník nezvládne vše přečíst
  - bibliografické údaje, kartotéky, knihovnické automatizované systémy, katalogizace, rešeršní služby (Lístkový lexikální archiv, 8,5 mil. lístků)
  - strojové zpracování **vědecko-technických** informací
  - KWIC – konkordance a jejich frekvence
- klasifikace textů
  - převedení věcného katalogu, nalezení literatury týkající se určitého tématu
  - třídění vědních disciplín – bibliografický klasifikační systém (v té době už 350)
  - desetinné třídění (čísla, silná hierarchie) – mezinárodní platnost
  - fasetové třídění (písmena, slabá hierarchie)

Strojová lingvistika

- **uložení dat** – ***machine-readable form*** (MRF)
  - identifikace znaků a jejich interpretace (děrné štítky, magnetické pásky), vytváření databází
- **ASCII** (*American Standard Code for Information Interchange*), 1960
  - ASA (*American Standard Association*), 1. vydání 1963
  - 128 znaků pro angličtinu, každý znak má binární hodnotu
- **OCR** (*Optical Character Recognition*)
  - pro češtinu špatné rozeznávání, nepodporovalo české znaky a některé fonty
  - **Ray Kurzweil**
    - původně pomůcka pro nevidomé – rozpoznávání znaků + syntéza řeči, 1976 představeno na konferenci, 1978 komerční prodej
    - 1980 prodáno firmě Xerox – převod tištěného textu do počítačově čitelné podoby

Strojová lingvistika

- **opětovné nalezení dat** – definování jazykových jednotek, jejich označení, segmentace
- **informace o obsahu**
- analýza textu, získávání informací (Information Retrieval)
  - indexování, klíčová slova, sémantické faktory (A. Kent, J. W. Perry – *Information Systems in Documentation*, 1957, Cleveland)
  - automatické referování, sumarizace textu (většinou statistické, výběr nejfrekventovanějších termínů z oblasti)
  - sestavování rejstříků

Strojová lingvistika

- Information Retrieval – Vannevar Bush (1890–1974), americký vědec, elektroinženýr, vizionář, státní úředník, vystudoval současně bc. i mgr. stupeň na Tufts Colege, doktorát současně z MIT i Harvardu
- pracoval na MIT, viceprezident, děkan, jeho student Claud Shannon, návrh digitálních okruhů
- analogový počítač pro diferenciální výpočty (až 18 proměnných)
- 1922 firma Raytheon (dnes Raytheon Company – jeden z největších dodavatelů zbraní a vojenské techniky)
- za války předseda *National Defense Research Committee* (NDRC), řídil činnost 6 tis. vědců za války v Americe, poradce prezidenta
- projekt Manhattan – inicioval, podporoval a zajistil mu prioritu
- *As We May Think*, 1945, nové formy encyklopedií (WWW, Internet, Wikipedia)
- Memex – prohlížeč mikrofilmů, ne indexování (na principu hypertextu, myš), jen vize

Strojová lingvistika

- komunikace člověka s počítačem, dialogové systémy, Question Answering Systems
- ELIZA – 1964–66 Joseph Weizenbaum, MIT
  - simulace dialogu pacienta s psychiatrem (klíčová slova), dnes *chatterbots*
  - znalostní báze (*knowledge base*)
  - Turingův test (Imitation Game)
  - vliv na RPG a filmy (G. Lucas – THX-1138), Siri
- ***word-processing*** – editory, psaní textů, práce s texty a slovy, korektory překlepů, dělení slov

Strojová lingvistika

70.–80. léta ČSSR v rámci VTEI (Informační soustava vědeckých, technických a ekonomických informací, zrušena po r. 1989)

- MOZAIKA – Zdeněk Kirschner, MFF UK, první experimenty 1977
  - automatická extrakce terminologie
  - podrobný popis výskytu relevantních termínů (segmentace, koncovky, sufixy, lemmatizace, omezení oborové, postavení ve větě, postavení v textu – váhy)
  - počítač EC 1040, 900–1000 slov/min (IBM 10 x rychlejší)
  - uživatel zadává indexy (klíčová slova)
  - vstup – psací stroj připojený k počítači, děrný štítek
  - výstup – tisk nebo obrazovka
  - možné využití také pro anotace, abstrakty (spíše seznam indexů)

Strojová lingvistika

- SEMAN – (pod VTEI) univerzální sémantický analyzátor, 80. léta
- Vladimír Smetáček (*1939), informatik, doc. pedagogiky, autor veršů pro děti, překladatel z ruštiny
- přirozený jazyk + sémantické rysy (sémy)
- (poloautomatizovaná) tvorba tezauru
- automatizovaná extrakce klíčových slov
- SEMAN po r. 1989 ztracen
- *Lidé a informace*, 1981
- *SEMAN – experimentální automatizovaný nástroj obsahové analýzy textů v přirozeném jazyce*, 1982