

PLIN041 Vývoj počítačové lingvistiky

Další aplikace matematiky v lingvistice
Teorie komunikace a teorie informace

Mgr. Dana Hlaváčková, Ph.D.

40.–70. léta 20. století

Glottochronologie

- nová jazykovědná disciplína, lexikální statistika (lexikostatistika), na poč. 50. let v USA
- v r. 1950 ji navrhl a popsal **Morris Swadesh**, americký lingvista, historicko-srovnávací lingvistika (1909–1967) – indiánské a eskymácké jazyky a dialekty
- inspirace radiokarbonovou metodou v chemii – rozpad radioaktivních jader uhlíku (poločas rozpadu)
- zjišťuje se původ jazyka a **doba rozpadu jazyka** na dva či více moderních jazyků
- příbuzenské jazykové vztahy se měří na základě změn v **základní slovní zásobě**

Glottochronologie

- jádro slovní zásoby, cca 200 slov označujících základní skutečnosti – *matka, otec, muž, žena, zvíře, malý, velký atd.* (*Swadesh list*)
- u dvou různých jazyků se porovnává 100 základních výrazů a měří se jejich shoda či rozrůzněnost v průběhu času
- na základě procenta shodných a různých dvojic se stanovuje **index rychlosti**, s jakou slova mizí z jádra slovní zásoby
- čas, kdy došlo k rozrůznění slovní zásoby – **časová hloubka** (podíl logaritmu procenta shodných dvojic a indexu rychlosti)

Glottochronologie

- pozitivní ohlasy u jazyků s mladší historií
- kritika indoevropéistů – jazyky s dlouhou historií, nesoulad výsledků
 - subjektivní výběr jádra slovní zásob
 - rozpad jádra neprobíhá konstantní rychlostí
 - podíl externích vlivů – převratná období, styk s jinými jazyky apod.
- použili **M. Čejka** a **A. Lamprecht** – rozpad praslovanské jednoty (větev jižní, východní a západní) – 8.–11. st. (s vrcholem ve st. 10.), ověřovali i tradičními metodami

Stylometrie

- frekvenční výzkumy v oblasti stylistiky – frekvence slov a gramatických kategorií v jednotlivých stylistických rovinách
- statistické charakteristiky stylu jednotlivých autorů – **určování autorství**
- typické a unikátní znaky autora, lze vyčíslit – otisk autora, **stylom**
- dnes s využitím strojového učení
- spory o autorství – Shakespeare, Jan Neruda, Rukopis královédvorský a zelenohorský

Teorie komunikace a informace

- matematika – první počítače, kybernetika, strojový překlad
- 40./50. léta nové vědní obory v matematice, výzkum přenosu informace, souvislost se vznikem kybernetiky
- **Claude Elwood Shannon** (angl. matematik), **Warren Weaver** (am. matematik, fyzik) – **The Mathematical Theory of Communication**, 1949, určeno matematikům
- **Charles Francis Hockett** (am. strukturalista) – **Review of Shannon & Weaver**, Language, 1953, recenze, přiblížil dílo lingvistům
- **Norbert Wiener** – zformuloval teorii informace nezávisle na Shannonovi a Weaverovi

Norbert Wiener

- 1894–1964, americký matematik a filozof, 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, vystudoval filozofii, ale disertace (v 17 letech) souvisela s matematickou logikou, Harvard (zoologie), Cambridge
- učil filozofii na Harvardu a matematiku na MIT, pracoval v oblasti balistiky
- teorie pravděpodobnosti, náhodné procesy a šum
- u studentů znám chabým způsobem přednášení, vtipy a roztržitostí
- dodnes je udělována Wienerova cena za aplikovanou matematiku

Norbert Wiener

- snažil se vstoupit do armády za 1. sv. v. (odmítán kvůli slabému zraku), přijat až na konci války jako prostý voják
- účastnil se prací v oblasti balistiky
- mezi válkami se oženil a měl dvě dcery
- za 2. sv. v. – střely na velkou vzdálenost
 - automat (servomechanismus)
 - 1) zasáhnout cíl, 2) odpovědět na otázku (zpětná vazba)
- na konci války – 1. radarem řízená střela (navádění během letu), pak se věnoval automatům (informace si pamatují)
- zformuloval teorii informace nezávisle na Shanonovi a Weaverovi
- po válce – kybernetika

Norbert Wiener

- kybernetika zkoumá stroje i živé organismy (pomezí disciplína)
- éra kybernetických strojů, počítače, analogie s lidským mozkiem (zpětná vazba na podněty z okolí u živých organismů i strojů)
- vynutila si teorii informace
- informatika, umělá inteligence, neuronové sítě
- na východě kybernetika nejdříve buržoazní pavěda – přijata na konci 50. let

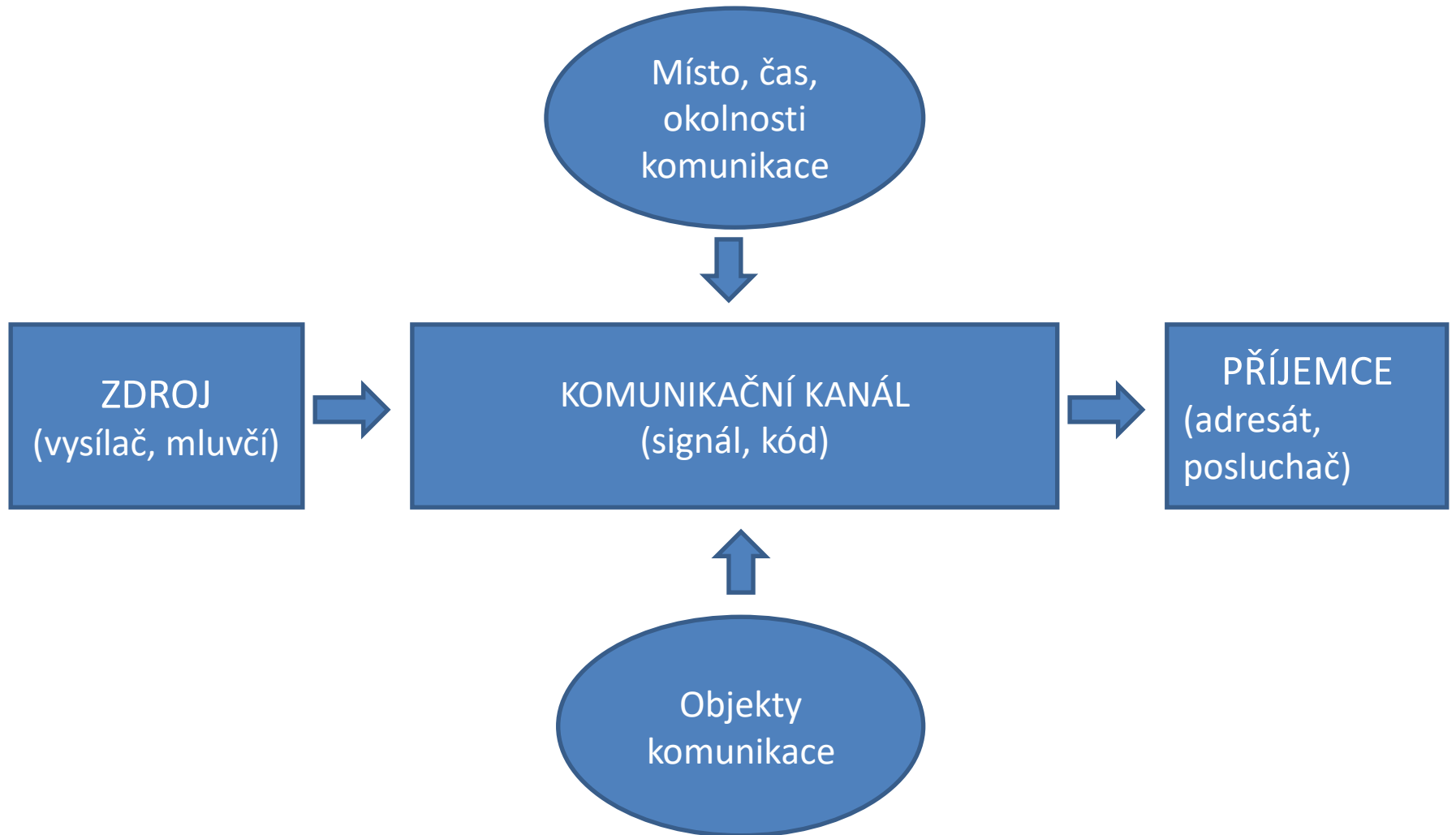
Claude Elwood Shannon

- americký matematik, elektroinženýr, kryptograf, „otec informačního věku“ (1916–2001)
- již v dětství nadání na matematiku a elektrotechniku (dálkově řízený model člunu, bezdrátový telegraf)
- MIT (návrh logických obvodů), stáž na Princetonu (setkání s významnými vědci – A. Einstein, J. von Neumann)

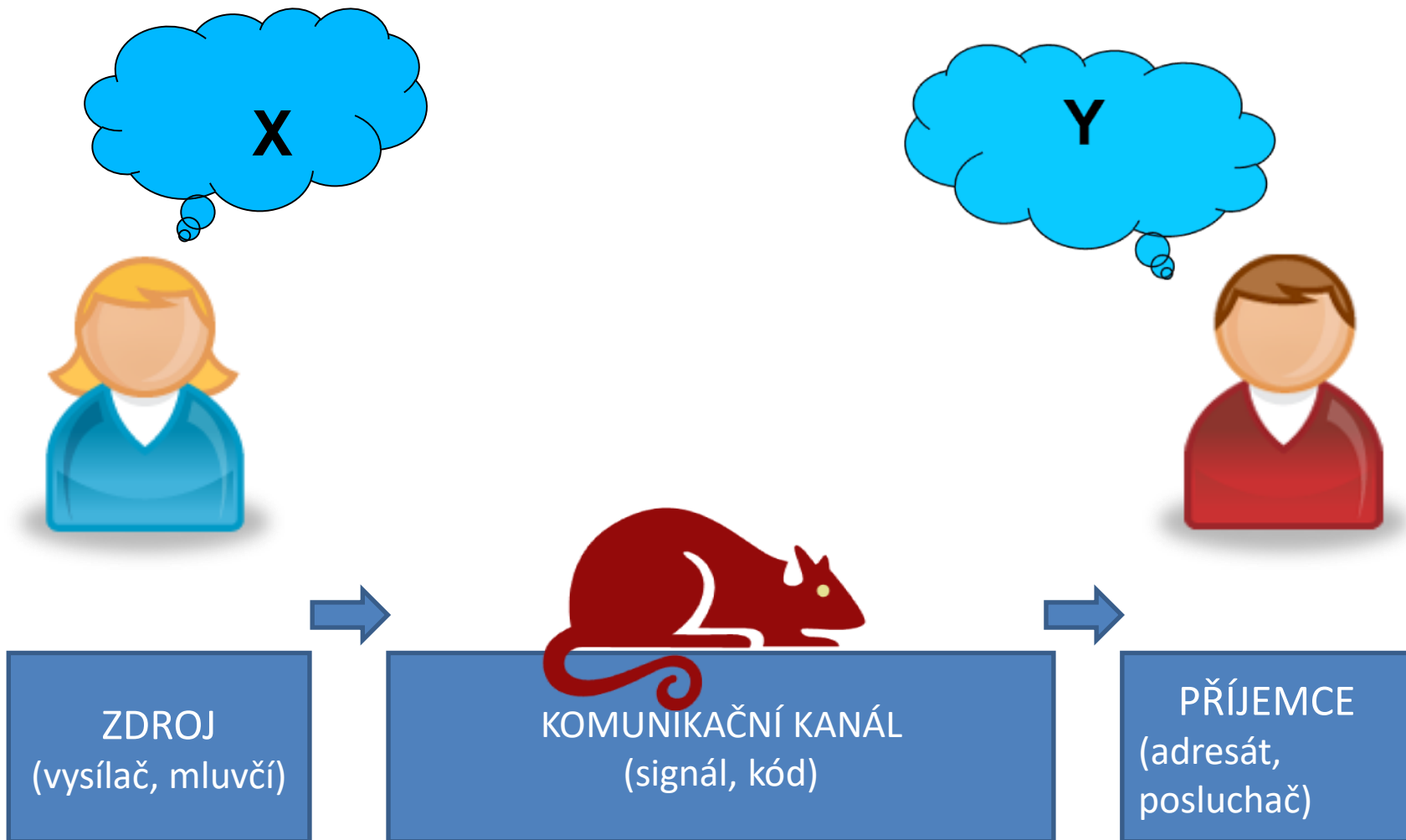
Claude Elwood Shannon

- 2. sv. v. – Bellovy laboratoře (zal. 1880 Alexander Graham Bell, dnes v New Jersey, vlastní Nokia)
- kryptografie, setkání s A. Turingem, přenos signálu a šumu – vyústilo v teorii informace
- **Prediction and Entropy of Printed English**, 1951
- vynálezy, např. Shannonova myš (učící se mechanismus, počátek UI)

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

Teorie informace

- nulová entropie = **redundance**, spolehlivost přenosu x šum
- polemika, výhrady
 - vztah entropie a frekvence (nižší frekvence vyšší entropie)
 - míra informace je individuální – zkušenost, vzdělání, věk
- jednotka množství informace – **bit** (binary digit), binární opozice 0/1, binarismus v jazykovědě (již ve strukturalismu)

Teorie informace

- teorie informace – kybernetika – strojová lingvistika – strojový překlad
- český sborník **Teorie informace a jazykověda**, 1964 (překlady zásadních článků z této oblasti)
- **Roland Lvovič Dobrušin** (ruský matematik) – **Matematické metody v lingvistice**, 1961 – využití matematických metod pro popis lingvistických jevů, zdokonalení strojového překladu
- **Warren Plath** (americký lingvista a matematik, Harvard) – **Matematická lingvistika**, 1961 – přehled dosavadního vývoje, statistické metody při určování autorství a příbuznosti jazyků

Teorie informace

- **C. E. Shannon** – **Predikace a entropie tištěné angličtiny**, 1951, metoda odhadu entropie a redundance, využití teorie informace ve zpracování přirozeného jazyka
- **V. V. Ivanov, S. K. Šaumjan** (přední sovětský strukturalista) – **Lingvistické problémy kybernetiky a strukturní lingvistika**, 1961, *Kibernetiku na službu kommunizmu*

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, na ideálním umělém jazyce a na angličtině; 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