PLIN041 Vývoj počítačové lingvistiky

Aplikace matematiky v lingvistice
Teorie komunikace a teorie informace

Mgr. Dana Hlaváčková, Ph.D.

40.–70. léta 20. století

Glottochronologie

- nová jazykovědná disciplína, též 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 indoevropeistů – 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

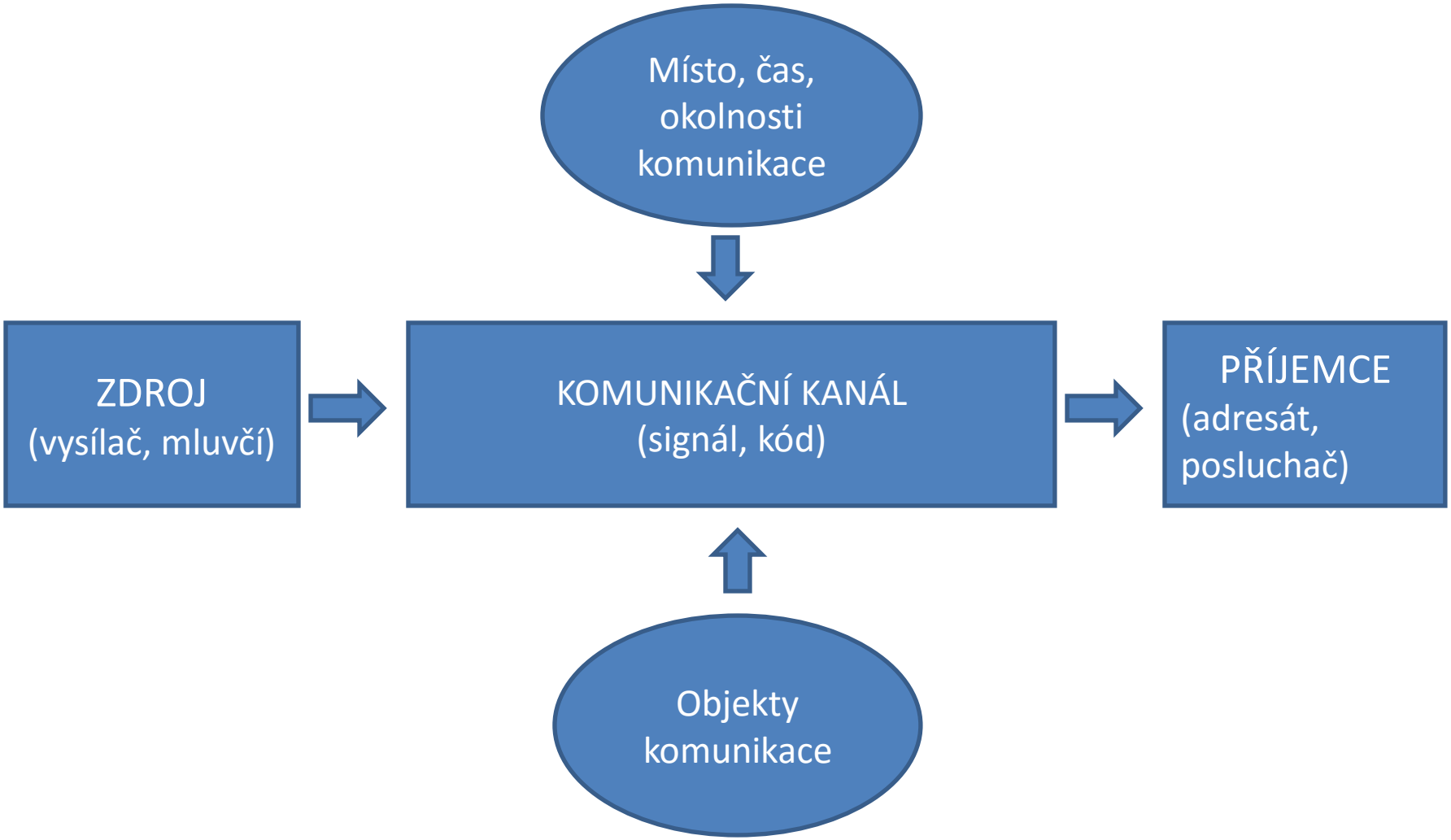
Stylometrie

- 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

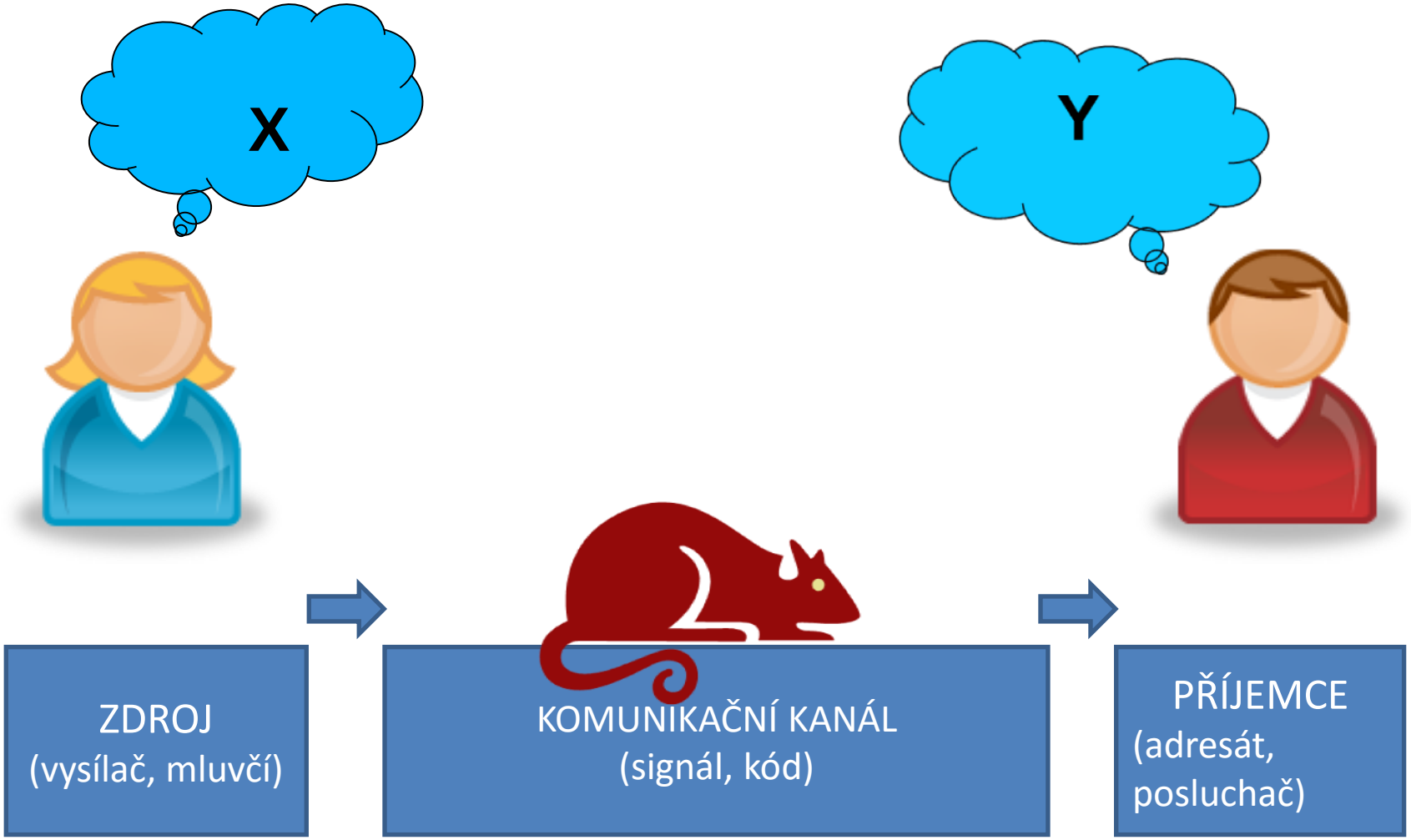
Teorie komunikace a informace

- 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 & Weawer, Language, 1953, recenze, přiblížil dílo lingvistům
- 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
- 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

- 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, metoda odhadu entropie a redundance, využití teorie informace ve zpracování přirozeného jazyka

Teorie informace

- Benoît Mandelbrot – Komunikace a formální strukura 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