PLIN059 Proseminář z počítačové lingvistiky – úvod

Mgr. Dana Hlaváčková, Ph.D.

hlavacko@phil.muni.cz

Ústav českého jazyka FF MU

A. Nováka 1, budova D

Formality

- účast povinná, možná 1 neomluvená absence
- jinak omluvenky do IS
- 1 vnitrosemestrální test
- 1 závěrečný test

Co je to počítačová lingvistika?

- Matematická lingvistika, Kvantitativní lingvistika, Komputační lingvistika, Počítačové zpracování přirozeného jazyka, Jazykové inženýrství
- Computational Linguistics, Mathematical Linguistics, Natural Language Processing (NLP)

- **průnik mezi lingvistikou, informatikou a matematikou**
- součást oblasti Digital Humanities

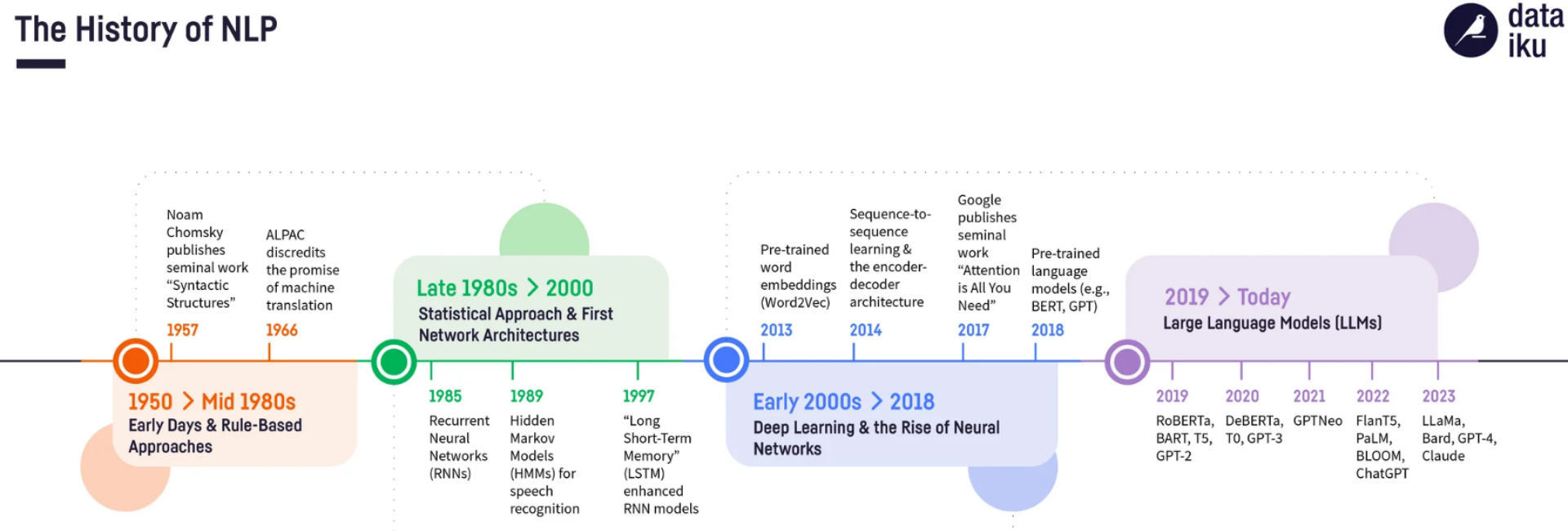
- **přirozený jazyk + počítačové zpracování**
- **detailní analýza jazyka a jeho formální popis**
- **strojové učení, neuronové sítě**
- **velké jazykové modely**

Co počítačová lingvistika dělá?

- výsledkem jsou denně používané **aplikace**, např.:
  - korektor překlepů
  - korektor gramatiky
  - vyhledávání na webu
  - prediktivní psaní
  - překladače jazyků
  - online slovníky
  - syntéza řeči
  - chatboty

Jak to počítačová lingvistika dělá

- pravidlový přístup (rule-based)
- stochastický přístup (statistika, pravděpodobnost)
- strojové učení, neuronové sítě (deep learning, neural networks)
- velké jazykové modely (LLM)



The Evolution of NLP (& How Dataiku Can Help)

Co počítačová lingvistika poskytuje?

- urychlení a zefektivnění práce lingvisty, zpracování velkých dat
- ověřování existujících teorií a hypotéz
- objevení nového jazykového jevu, zákonitosti
- počítačový lingvista ví, co a jak může použít
- umí sám nástroj vytvořit
- autorská práva a přístupy k nástrojům
  - https://1url.cz/81FiC
  - veřejně dostupné, příp. dostupné na MU
  - vlastní přístup (registrace)

Pár obecných zásad

- proč to chceme? (cíl, účel, uživatel)
- jak toho dosáhneme? (efektivita)
- uživatelská přívětivost, uživatelská zkušenost, User Experience
- maximum automatizace – minimum ruční práce (při vytváření i používání)
- zpracování velkého objemu dat
- univerzálnost (široká množina vstupů, spojování více nástrojů do jednoho)
- nezávislost na jednotlivých lingvistických teoriích
- dříve desktopové, dnes webové aplikace
- při zpracování i používání je nutná PŘESNOST

Obsah kurzu

- **počítačová lexikografie** – DEBDict, DEBWrite, ASSČ, Vokabulář webový a další
- **jazykové korpusy** – KonText, Sketch Engine
- **morfologická analýza** – Ajka, Majka, Morče, MorphoDiTa (atributivní a poziční systém)
- **derivační** rozhraní – Morfio, DeriNet
- **syntaktická** analýza – Synt, Set, PDT (stromové banky)
- **sémantická** analýza – WordNet, FrameNet, VerbNet
- **valenční databáze** – Vallex, VerbaLex
- **rozpoznávání a syntéza řeči**

Příbuzná pracoviště

- **Centrum zpracování přirozeného jazyka** FI MU Brno – http://nlp.fi.muni.cz/
- **Ústav formální a aplikované lingvistiky** MFF UK Praha – http://ufal.mff.cuni.cz
- **Ústav teoretické a komputační lingvistiky** FF UK Praha – http://utkl.ff.cuni.cz
- **Ústav Českého národního korpusu** FF UK Praha – http://www.korpus.cz
- **Ústav pro jazyk český** AV ČR – http://www.ujc.cas.cz

Příbuzná pracoviště

- **Fakulta informačních technologií** VUT Brno – http://www.fit.vutbr.cz
- **Katedra informatiky a výpočetní techniky** – http://www.kiv.zcu.cz, **Katedra kybernetiky** http://www.kky.zcu.cz FAV ZČU Plzeň
- **Ústav informačních technologií a elektroniky** FM TU Liberec – http://www.fm.tul.cz
- **Slovenský národný korpus**, JÚĽŠ SAV Bratislava – http://korpus.juls.savba.sk/

Ukázky

- Internetová jazyková příručka http://prirucka.ujc.cas.cz
- Slovo v kostce https://www.korpus.cz/slovo-v-kostce/
- Opravidlo https://www.opravidlo.cz/