Korpusová lingvistika

CJBB85 Počítačové nástroje pro češtinu

Mgr. Jakub Machura

415795@mail.muni.cz

Co nás čeká?

- nástroje běžící na pražské platformě:
  - **SyD**
  - **Morfio**
  - **Kwords**
  - **Treq**
- brněnský korpusový manažer **Sketch Engine**
  - **Word Sketch**
  - **Tezaurus**
  - **N-gramy**
  - **Skell**

- https://app.sketchengine.eu/
- http://ske.fi.muni.cz
- http://korpus.cz

SyD

- https://syd.korpus.cz/
- nástroj pro korpusový průzkum variant
- synchronní a diachronní část

Morfio

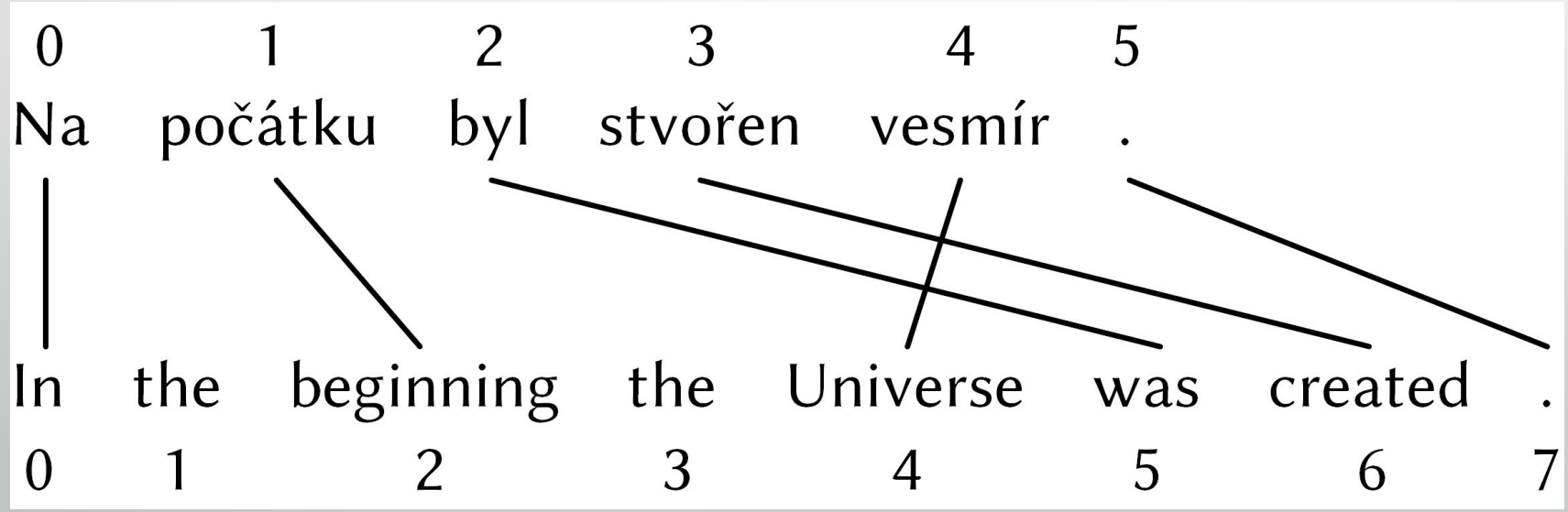
- https://morfio.korpus.cz/
- slovotvorný výzkum, zejména derivace
- více informací v hodině věnované slovotvorbě

KWords

- https://kwords.korpus.cz/
- nástroj pro identifikaci klíčových slov
- součást projektu *A Needle in a Haystack*
- zkoumaný text je porovnáván s korpusem (referenční text)

Treq

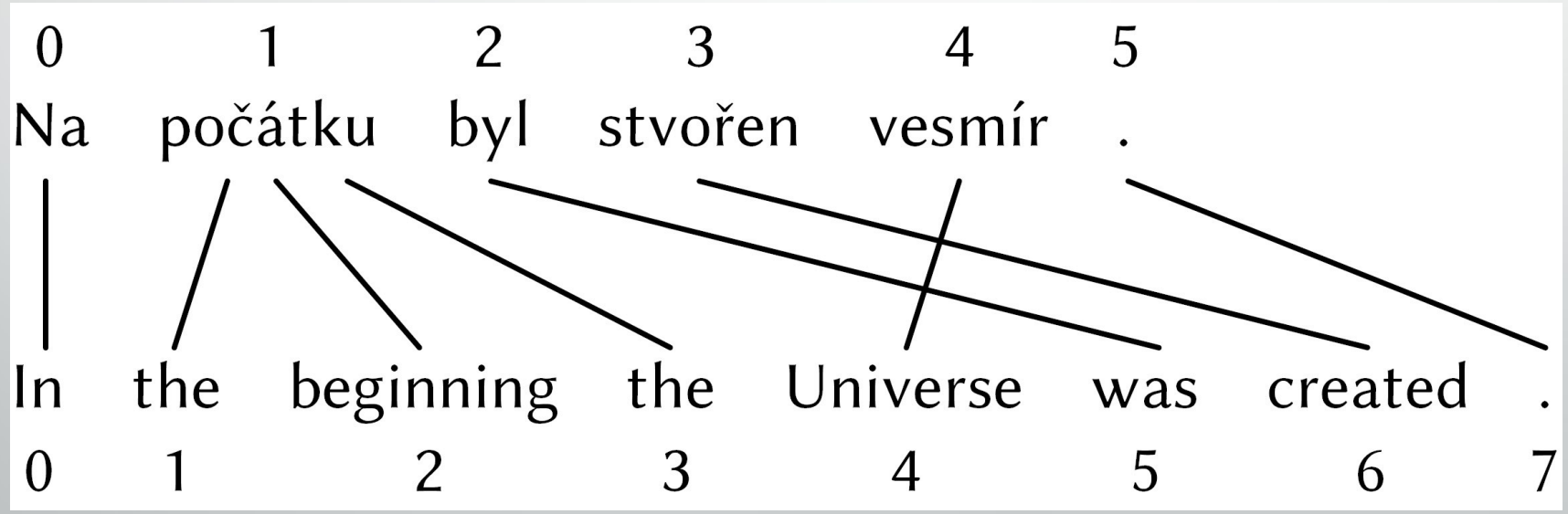
- https://treq.korpus.cz/
- databáze překladových ekvivalentů
- vytvořeno automaticky na základě dat z paralelního korpusu InterCorp.



Převzato z
http://wiki.korpus.cz/lib/exe/fetch.php/manualy:carky_gdfa.jpg?cache=

Treq

- https://treq.korpus.cz/
- databáze překladových ekvivalentů
- vytvořeno automaticky na základě dat z paralelního korpusu InterCorp.



Převzato z
http://wiki.korpus.cz/lib/exe/fetch.php/manualy:carky_gdfa.jpg?cache=

Sketch Engine

- komerční korpusový manažer
- Lexical Computing
- 500 korpusů (převážně webových) ve více než 90 jazycích
- atributivní značkovací systém
- tagset: https://nlp.fi.muni.cz/projekty/ajka/tags.pdf
- https://app.sketchengine.eu/

Sketch Engine – Word Sketches

- tzv. slovní profily
- sdružují kolokace slov na základě gramatických relací (podmět, přísudek, atribut, …)
- hodnoty frekvence a skóre
- https://old.sketchengine.co.uk/corpus/wsdef?corpname=preloaded/cstenten17_mj2

Sketch Engine: Tezaurus

- obecně: „Ucelený obraz lexikálního systému (popř. jeho části) v slovníkové podobě, organizovaný pojmově, významově, věcně; vystavěný na principu onomaziologickém (tj. od významu k formě).“ (Čermák, Hladká. NESČ, https://www.czechency.org/slovnik/TEZAURUS)
- využití Word Sketches a jejich následného porovnání

Sketch Engine: n-gramy

- sled po sobě jdoucích položek z dané posloupnosti
- unigram, bigram, trigram…
- kolokace vs. n-gram

The office building was demolished yesterday.

- 5 bigramů a 2 kolokace

office building

to demolish a building

Sketch Engine: SkELL

- https://csskell.sketchengine.co.uk/run.cgi/skell

- webová aplikace pro učení jazyka na základě dat z korpusu