# CJBB105 Úvod do korpusové lingvistiky

**přednáška pro oba cykly studia**

Ukončení: kolokvium (písemka ověřující znalosti získané na přednáškách nebo studiem příslušné literatury)

út.: 11.40-13.10 A33

# dnes

- harmonogram přednášek
- studijní literatura
- širší perspektivy oboru

# Osnova přednášky:

- 1. Úvod – literatura 26.9.
- 2. Historie KL 3.10
- 3. Co to je korpus a co v něm můžeme najít 10.10.
- 4. Kvantitativní data 17.10.
- 5. Využití korpusu pro lingvistická bádání 24.10.
- 6. Korpusy a počítačová lingvistika 31.10.
- 7. Morfologická analýza a tagování korpusu 7.11.
- 8. Samostatné studium ČNK – www 14.11.
- 9. Korpusová lingvistika u nás – Český národní korpus 21.11.
- 10. Korpusy na MU 28.11.
- 11. Korpusové manažery (BONITO - ČNK) 5.12.
- 12. Kolokvium I. - předtermín 12.12.
- 13. Kolokvium I. - I. termín 19.12.

# Studijní literatura
# učebnice

- Barnbrook G. (1996): Language and Computers. Edinburgh University Press, Edinburgh.
- McEnery A., Wilson A. (1996): Corpus Linguistics. Edinburgh University Press, Edinburgh.
- Šulc M.: Korpusová lingvistika. První vstup. Praha : Karolinum. 1999.

# Monografie / sborníky

- Blatná R., Petkevič, V. (eds.) (2005): Jazyky a jazykověda. Sborník k 65. narozeninám prof. Františka Čermáka. Praha : FF UK – ÚČNK, s.
- Čermák F., Klímová J., Petkevič V. (eds.) (2000): Studie z korpusové lingvistiky , Praha: FF UK.
- Čermák F, Blatná R. (eds.) (1995): Manuál lexikografie. Jinočany : H&H.

# Články

- Čermák, F.: Jazykový korpus: Prostředek a zdroj poznání. SaS, 56, 1995, s. 119-140.
- Čermák F., Králík J., Kučera K. (1997): Recepce současné češtiny a reprezentativnost korpusu (Výsledky a některé souvislosti jedné orientační sondy na pozadí budování Českého národního korpusu). SaS, 58, 2, s. 118-124.
- Čermák F. (1999): Oxfordská lexikografie přechází také plně na korpus. Slovo a slovesnost, 60, s. 136-141.

# Encyklopedie

- Karlík P., Nekula M., Pleskalová J. (eds.) (2002): Encyklopedický slovník češtiny. Praha : Nakladatelství Lidové noviny.

# Nejdůležitější www

- http://ucnk.ff.cuni.cz/
- http://www.athel.com/corpus.html
- http://www.tei-c.org/
- http://nlp.fi.muni.cz/

# Korpusová lingvistika – širší souvislosti

(lingvistika – matematika – umělá inteligence – informatika)

- komputační lingvistika –(NLP, language ingeneering)
- kvantitativní lingvistika
- algebraická lingvistika
- korpusová lingvistika

# Počátky matematické lingvistiky

- Strukturalismus (PLK)
- Kvantitativní lingvistika
- Omezení v 50. letech
- 60. léta - překladový sborník *Teorie informace a jazykověda* (1964)

# Kvantitativní lingvistika

- FSČ (1961)
- Oddělení kvantitativní lingvistiky ÚJČ
- 70. léta – první počítačově čitelný korpus (540 000 slovních výskytů)
- 80. léta – řada FS (M. Těšitelová)
- 1994 založení Journal of Quantitative Linguistics - International Quantitative Linguistics Association (IQLA)

# Matematická lingvistika

- Konec 50. let -Oddělení teorie strojového překladu FF UK
- *1964 Cesty moderní jazykovědy* (*Jazykověda a automatizace*)
- FGP (FGD)
- Petr Sgall (1967) *Generativní popis jazyka a česká deklinace*

# Strojový překlad

- První pokus (SAPO – VÚMS) leden 1960
- *Učíme stroje česky* (Sgall, Hajičová, Piťha, 1986).
- APAČ (1977-1986)
- MATRACE (1990-1992)

# Počátky KL u nás

- 1992 - *Počítačový fond češtiny* (PFČ )
- 1993-95 Počítačový korpus českých psaných textů
- 1995 ÚČNK
- Čeština ve věku počítačů (Komplexní projekt GAČR) 1996-2001

# Grantové projekty

- Počítačový korpus českého jazyka (Posílení výzkumu na vysokých školách, MŠMT ČR), 1996-2000
- Výzkumný záměr MŠMT "Český národní korpus a korpusy dalších jazyků" (1999-2004)

# Grantové projekty

- Korpus českých psaných textů (V. Petkevič, Grantová agentura České republiky)
- Programové nástroje pro počítačové zpracování českých textů (J. Peregrin, Grantová agentura České republiky)
- Česká frazeologie, její výzkum a lexikografické zpracování (F. Čermák, GAUK)

# Grantové projekty

- Korpus mluvené češtiny v počítačovém zpracování (F. Čermák, GAUK)
- Elektronizace postupů diachronní lexikografie (P. Nejedlý, R. Blatná, Grantová agentura České republiky)

# Grantové projekty

- **Velké jazykové korpusy a jejich automatická analýza**, GAČR (2003-2005)
- Výzkumný záměr MŠMT **Český národní korpus a korpusy dalších jazyků**, VZ MSM 0021620823, (2005-2011)

# Grantové projekty FI MU

- http://nlp.fi.muni.cz/nlp/aisa/NlpCz/Grantove_projekty.html

# Grantové projekty FF MU

- ***Současná soukromá korespondence. Vytvoření databáze a zpracování vybraných jevů z pohledu lexikologicko-lexikografického a dialektologického***

# Ústav pro jazyk český AVČR

- ***Možnosti a meze gramatiky češtiny ve světle Českého národního korpusu***
- ***(Konference: Korpus jako zdroj dat o češtině 4. - 6. listopadu 2004 +***
- ***Sborník: Karlík, P. (red.): Korpus jako zdroj dat o češtině, Brno : FF MU. 2005.***

# FF UK

- Ústav formální a aplikované lingvistiky MFF UK http://ckl.ms.mff.cuni.cz/ufal/
- Ústav teoretické a komputační lingvistiky FF UK http://utkl.ff.cuni.cz/

# PZK - PDT

- Prague Dependency Treebank (PDT) – korpus anotovaný na dvou syntaktických úrovních zahrnujících údaje o aktuálním členění a hlavních typech koreference
- http://ufal.mff.cuni.cz/pdt2.0/

# Další univerzitní pracoviště v ČR

- ***Fakulta aplikovaných věd - FAV ZU http://www.kky.zcu.cz/,***
- ***Ústav informatiky Filozoficko-přírodovědecké fakulty Slezské univerzity v Opavě***
- *Katedra elektroniky a zpracování signálů Technické univerzity v Liberci*

# SBORNÍKY specializované na ML

- PBML (The Prague Bulletin of Mathematical Linguistics http://ufal.mff.cuni.cz/?a=pbml).
- PSML (The Prague Studies in Mathematical Linguistics 1-10, 1964-1990).

# Články

- Jednotlivé články k oboru je možno najít v lingvisticky orientovaných sbornících a časopisech (Slovo a slovesnost, Naše řeč aj.) a dále např. v časopisech Čs. Informatika, Kybernetika, Czechoslovak Mathematical Journal.

# Přehled slovníků

- Jelínek Jaroslav, Bečka Josef, V., Těšitelová Marie (1961): Frekvence slov, slovních druhů a tvarů v českém jazyce. Praha : SPN.
- Slavíčková Eleonora (1975): Retrográdní morfematický slovník češtiny s připojenými inventárními slovníky českých morfémů kořenových, prefixálních a sufixálních. Praha: Academia.

# slovníky

- Těšitelová Marie (1980): Frekvenční slovník současné české publicistiky, Praha : Ústav pro jazyk český ČSAV.
- Těšitelová Marie (1980): Frekvenční slovník současné české administrativy, Praha : Ústav pro jazyk český ČSAV.
- Těšitelová Marie (1983): Frekvenční slovník jazyka věcného stylu, Praha : Ústav pro jazyk český ČSAV.
- Těšitelová Marie (1983): Frekvenční slovník současné odborné češtiny, Praha : Ústav pro jazyk český ČSAV.

# slovníky

- Králík Jan, Těšitelová Marie (1986): Retrográdní slovník současné češtiny. Praha: Academia.
- Pala Karel, Všianský Jan (1994): Slovník českých synonym. Praha: Nakladatelství Lidových Novin. (2. vyd. 2001.)
- Čermák František, Křen Michal (2004): Frekvenční slovník češtiny. Praha: NLN. + 1 CD-ROM.

# Bibliografická poznámka

- ACL - mezinárodní vědecké profesní sdružení lidí, kteří se zabývají problémy NLP.
- Digital Archive of Research Papers in Computational Linguistics http://acl.ldc.upenn.edu/.