

# Informace ke státní závěrečné zkoušce programu FF N-PLIN Počítačová lingvistika pro studenty imatrikulované od podzimu 2019 včetně

## PRO VEDLEJŠÍ STUDIJNÍ PLÁN

### PREAMBULE

Vedlejší studijní plán neobsahuje obhajobu práce.

### ÚSTNÍ ZKOUŠKA

Otázky jsou vybírány na základě tematických okruhů, které mají studentům sloužit jako opora pro přípravu k SZZ. Vycházejí z látky probrané v povinných a profilových kurzech navazujícího magisterského studijního programu Počítačová lingvistika na FF MU.

### OBECNÉ PŘÍRUČKY DOPORUČENÉ PRO PŘÍPRAVU KE SZZ

K části orientované na **český jazyk**:

- Karlík, P. – Nekula, M. – Rusínová, Z. (eds.). *Příruční mluvnice češtiny*. Nakladatelství Lidové noviny: Praha, 1995.
- Karlík, P. – Nekula, M. – Pleskalová, J. (eds.). *Nový encyklopedický slovník češtiny*. Nakladatelství Lidové noviny: Praha, 2016.

K části orientované na **počítačovou a korpusovou lingvistiku, informatiku**:

- Čermák, F. – Klímová, J. – Petkevič, V. (eds.) *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000.
- Osolsobě, K. *Matematická lingvistika*. In Krčmová, M. – Večerka, R. – Pleskalová, J. – Karlík, P. *Kapitoly z dějin české jazykovědné bohemistiky*. Praha: Academia, 2007, s. 447–466.
- Kocek, J. – Koprivová, M. – Kučera, K. (eds.) *Český národní korpus – úvod a příručka uživatele*. Praha: ÚČNK FF UK, 2000.
- Osolsobě, K. *Česká morfologie a korpusy*. 1. vyd. Praha: Karolinum, 2014.

**Literatura k jednotlivým okruhům** je zadána v sylabech povinných a povinně volitelných kurzů bakalářského studia v IS MU.

## **TEMATICKÉ OKRUHY KE SZZ**

Přehled okruhů z lingvistiky, matematiky, statistiky, informatiky, počítačového zpracování přirozeného jazyka a korpusové lingvistiky pro studenty navazujícího magisterského studia programu Počítačová lingvistika imatrikulované **od podzimu 2019** včetně (vedlejší studijní plán).

### **I. LINGVISTIKA (NA MATERIÁLU ČEŠTINY)**

#### **Čeština v komparativní perspektivě**

1. Subjekt, komplement, adjunkt.
2. Pro-drop parametr.
3. Konstrukce s reflexivem a jejich analýza.
4. Tranzitiva, ditranzitiva, neakuzativa, neergativa.
5. Slabičná struktura – slabičné konsonanty, extraslabičné konsonanty, alternace vokálů s nulou.

#### **Čeština v teorii a praxi**

1. Slovní druhy – kritéria klasifikace (morfologická, syntaktická, sémantická).
2. Substantiva – gramatické kategorie substantiv; deklinační paradigmata.
3. Adjektiva – typy deklinace (složená, jmenná, smíšená deklinace posesivních adjektiv).
4. Slovesa – gramatické kategorie sloves; finitní a nefinitní tvary, syntetické a analytické formy; konjugační paradigmata/slovesné třídy.
5. Větné členy – podmět, přísudek, přívlastek, předmět, příslovečné určení, doplněk (jak je lze poznat a jaké mají vlastnosti). Souvětí – souřadné a podřadné.
6. Slovní zásoba – struktura; vývojové tendence, neologismy.
7. Lexikografie – předmět zájmu; stavba slovníku, prezentace makrostruktury a mikrostruktury na vybraném slovníkovém díle; typologie slovníků, zastoupení jednotlivých typů v české lexikografii od počátku 20. století;

výkladový slovník – charakteristika, základní výkladové slovníky češtiny: PSJČ, SSJČ, SSČ; perspektivy české výkladové lexikografie.

8. Čeština v rámci slovanských jazyků, vývojové fáze češtiny.
9. Teritoriální stratifikace národního jazyka – nástin českých dialektů; nivelizační procesy, interdialekty a obecná čeština; teritoriální diference v užívání spisovné češtiny.
10. Norma, úzus, kodifikace – péče o spisovný jazyk; aktuální kodifikační příručky.

## **II. MATEMATIKA, INFORMATIKA, POČÍTAČOVÉ ZPRACOVÁNÍ JAZYKA**

### **Matematika a statistika**

1. Logika a její aplikace na přirozený jazyk.
2. Aplikace statistiky a pravděpodobnosti na zpracování jazyka.
3. Statistické jazykové modely (n-gramový a další) – konstrukce a využití, alternativy.
4. Vyhodnocování aplikací zpracování jazyka – přesnost, pokrytí, F-míry a zlatý standard, křížová validace.

### **Informatika a programování**

1. Reprezentace jazykových informací v počítači – kódování znaků, morfologická, syntaktická anotace, značkovací jazyky ((X)HTML, XML, DTD, ...).
2. Formální jazyky a jejich aplikace na zpracování přirozeného jazyka – regulární gramatiky a regulární výrazy, formální gramatiky a jejich využití v analýze morfologie a syntaxe.
3. Strojové učení a umělá inteligence – jaké problémy řeší a jaké algoritmy používá. Principy, techniky, vyhodnocování. Křížová validace.

### **Počítačové zpracování přirozeného jazyka**

1. Word sense disambiguation – definice, techniky a algoritmy, vyhodnocování, problémy.
2. Regulární jazyky a regulární výrazy, CQL.

3. Strojový překlad – pravidlové a statistické přístupy, existující systémy, vyhodnocování.
4. Morfologická analýza přirozeného jazyka a desambiguace – techniky a algoritmy, vyhodnocování, problémy.
5. Syntaktická analýza přirozeného jazyka, syntaktické stromy závislostní a složkové, algoritmy analýzy.
6. Budování a indexace korpusů, webové korpusy, paralelní korpusy. Statistické nástroje pro korpusy, jazykové modelování.
7. Počítačová lexikografie – systémy pro editaci slovníků, značkování slovníkového hesla.
8. Reprezentace znalostí, jazyková znalost, znalost o světě, encyklopedická znalost, common-sense.
9. Dialogové systémy – dialogové strategie, typy dialogů, značkovací jazyk pro popis dialogu, využití dialogových systémů, komunikační agenty (chatbots).

### **III. POČÍTAČOVÁ A KORPUSOVÁ LINGVISTIKA**

1. Počítačové zpracování jednotlivých rovin přirozeného jazyka z pohledu lingvistiky – morfologická, syntaktická, sémantická, slovtvorná.
2. Morfologický slovník jakožto součást automatického analyzátoru – zachycení gramatických významů v morfologickém slovníku, zachycení standardních a substandardních tvarů v morfologickém slovníku.
3. Automatické nástroje pro studium gramatiky budované nad jazykovými korpusy – konkrétní aplikace, využití složitějších dotazů v jazyce CQL pro studium gramatického systému jazyka.
4. Výběr vhodného korpusu pro řešení lingvistického problému – volně dostupné korpusy a jejich charakteristika, DIY korpusy, lingvistické příručky založené na korpusech.