Zpracování staré češtiny s novočeskými modely

Datum konání: 18. 10. 2022

Místo konání: D32

Název přednášky: Zpracování staré češtiny s novočeskými modely

Přednášející: RNDr. Daniel Zeman, Ph.D.

Počet účastníků: 20

Zpracovali: Blažková Eva

Biogram

Daniel Zeman působí na Ústavu formální a aplikované lingvistiky FF na Univerzitě Karlově. Mezi jeho hlavní výzkumné zájmy patří morfologie, syntax a strojový překlad. Zároveň pracuje na projektu Universal Dependencies.

Obsah zprávy

Universal Dependencies je mezinárodní projekt, který má hlavní cíl sjednotit značkování syntaxe pro všechny jazyky. Značkovací styl by měl být jednoduchý a jednotný, aby se stejným výrazem dalo hledat ve více jazycích. Jednotné značkování by dále mělo být rozšířené o specifika daného jazyka. Současně je v projektu zachyceno 130 různých jazyků, které ale nejsou reprezentovány ve stejné míře. Dominují jazyky indoevropské.

Universal Dependencies se dají využít, např. při výuce cizích jazyků (dostaneme konkrétní struktury vět, a to nám může ulehčit porozumění danému jazyku). Dalším využitím v rámci Digital Humanities jsou např. vzniklé databáze hliněných destiček akkadštiny. Můžou sloužit i k dokumentaci jazyků, které jsou v ohrožení, a i k použití na lingvistické typologie.

Zpracování staré češtiny

Na začátku stála myšlenka, zda by bylo možné udělat „treebank“ na textech ze 14.–15. století. Použita byla Drážďanská a Olomoucká bible, konkrétně Matoušovo evangelium. Tyto texty obsahují dohromady 44 tisíc slov. Původní představa byla, že by se vzal parser natrénovaný na podobných datech, nechaly by se texty zanalyzovat, ručně by se

opravila data, která by se přidala na trénovaní parseru a znovu by se spustil. Vyskytlo se však několik problémů. Jedním z nich je pravopis, který nemá v moderní češtině stejnou podobu. Dále obsahuje gramatické jevy, např. duál, aoristy, které v moderní češtině také nenalezneme.

Zpracování staré češtiny je prozatím na počátku, potýká se s problémem malého vzorku dat, která nejsou reprezentativní. Natrénování modelů bude stát ještě hodně úsilí.

Závěr

Přednáška mi připadala velmi zajímavá. Stará čeština a syntax patří mezi mé oblíbené zaměření oboru. Přínos vidím především v představení Universal Dependencies a ukázku jejich používání. Pokud by se podařilo natrénovat modely na staré češtině, myslím, že by takovou databázi ocenil nejeden student bohemistiky při studiu, ale i spousta odborníků.