

Práce navazuje na publikaci N-gram Text Categorization pánů Cavnara a Trenkla [2], rozšiřuje studium možnosti aplikace statistických metod na určení jazyka dokumentu jako jednodušší kategorie. Zde navrhovaný program čítá několik předností, mimo jiné modularita přiloženého programu a realizace nad jednou zvolenou univerzální znakovou sadou. Jako vylepšení původní práce je použito automatického filtrování a extrakce textu z internetových stránek pro průběžné zpřesňování naučených dat. Studujeme vývoj přesnosti určování jazyka dokumentu se zaměřením na krátké textové úseky. Tento aspekt jsme nenalezli v dostupné literatuře, jelikož kategorizace textu se ve většině případů zabývá dostatečně dlouhými vstupy. V závěru diskutujeme o možnosti použít naučená data a přístup pro označení a určení případných z jiného jazyka vnesených slov nebo krátkých frází. Konkrétně zkoumáme aplikaci na pražský závislostní korpus [8], kde tyto cizojazyčné fráze nejsou určeny, jen vymezeny.