

Tato práce navazuje na implementačně-výzkumný projekt Morče, jehož cílem bylo vytvoření co nejlepšího morfologického taggeru češtiny, založeného na skrytém Markovově modelu s průměrovaným perceptronem. Úspěšnost algoritmu závisí především na zvolené sadě rysů popisujících kontext, na jehož základě se značky vybírají. Práce stručně popisuje zvolený algoritmus a jeho implementaci. Její stěžejní část spočívá ve velké řadě provedených experimentů, které v rámci daných možností důkladně mapují možné sady rysů, jejich úspěšnosti a vztahy mezi nimi. Pro tento účel jsou definována pravidla, podle kterých se verze porovnávají. Využívá se pětinasobná crossvalidace a pro zjištění statistické významnosti výsledků je aplikován t-test. Při zahájení práce byla dána k dispozici nová data pro češtinu, takže veškeré experimenty se již prováděly nad daty z PDT 2.0. Vedlejším výsledkem práce je i statisticky významné zvýšení úspěšnosti taggeru, nicméně nejlepší tagger zřejmě překonán nebyl. Kromě ručního vývoje verzí byl projekt také upraven pro automatický vývoj, který byl v menším rozsahu proveden a popsán.