

Součástí úkolu jsou Python3 skripty `assignment3_1.py` a `assignment3_2.py`.

## Task 1

Použili sme Brillův tagger z knihovny NLTK. Ako počiatkový tagger sme použili jednoduchý unigramový tagger, ktorý si zapamätá najčastejšie značku pre každé slovo tréningových dát a pre neznáme slová vráti špeciálnu značku “None”. Použili sme počiatkovú množinu 24 vzorov<sup>1</sup> (template), ktoré pôvodne použil Brill. Pre angličtinu sme početpravidiel obmedzili<sup>2</sup> na 247, pre češtinu na 500. Tieto konkrétne hodnoty sme dostali skúšaním niekoľkých hodnôt. Všeobecne platí, že väčšie množstvo pravidiel môže mierne zlepšiť presnosť, ale výrazne spomaľuje.

## Výsledky

Angličtina	Přesnost	Čeština	Přesnost
1. stupeň	0.8909342163585638	1. stupeň	0.7743503905627291
2. stupeň	0.8890974014126724	2. stupeň	0.7887507306445614
3. stupeň	0.8747591522157996	3. stupeň	0.7618758256274769
4. stupeň	0.8935028395769291	4. stupeň	0.7919649224755012
5. stupeň	0.8925100950892275	5. stupeň	0.7880596223212392
Průměr	0.888160740931	Průměr	0.781000298326
$\sigma$	0.00686479392893	$\sigma$	0.0113145302014
Průměr triv.	0.865374314735	Průměr triv.	0.738888267001
$\sigma$ triv.	0.0143254860843	$\sigma$ triv.	0.0134321134672

Tabulka 1: Přesnost (accuracy) pro Brillův tagger pro oba jazyky.

## Task 2

### Supervizovaný HMM

Metóda	Čas	Přesnost
Žiadny pruning, add $\lambda$	7h 24m 22s	0.9252470471671056
Nevyhladený OOV model	1m 34s	0.9268375355252523
Vyhladený OOV model	5h 35m 13s	0.9280369201887727
OOV a obmedzené stavy	1h 12m 24s	0.927802257971997
Add $\lambda$ a obmedzené stavy	1h 23m 32s	0.9250123849503298
OOV model, pruning 10	1m 53s	0.9279847730294892
OOV model, pruning 10	3m 45s	0.9280369201887727

Tabulka 2: Přesnost (accuracy) pro Brillův tagger pro oba jazyky.

---

<sup>1</sup>NLTK má tieto vzory ako súčasť knižnice

<sup>2</sup>Limit sa pri týchto počtoch vždy naplní, čiže limit je rovnaký ako počet pravidiel.

Angličtina	Přesnost	Čeština	Přesnost
1. stupeň	0.9280369201887727	1. stupeň	0.791274775492853
2. stupeň	0.9227200458727552	2. stupeň	0.8096870184388119
3. stupeň	0.9277601208774032	3. stupeň	0.7636195508586526
4. stupeň	0.9170442118418997	4. stupeň	0.8034285110541747
5. stupeň	0.93424514784421	5. stupeň	0.8035752174835561
Průměr	0.92596128932500821	Průměr	0.80233748927646664
$\sigma$	0.00576517615271536	$\sigma$	0.0060165626151515

Tabulka 3: Přesnost pro oba jazyky určené cross-validací pro model z prořezávacím limitem 20. Pro češtinu se omezejeme pouze na stavy, které jsou v trénovacích datech, pro angličtinu používáme jako stavy všechny dvojice značek.

## Nesupervizovaný HMM — Baum-Welch

Pre túto časť sme sa rozhodli použiť ako stavy iba bigramy, ktoré sme videli v trénovacích dátach, pretože inak by bol počet stavov bol veľký.

Navyše sme sa rozhodli použiť pruningovú metódu, aby sme tréning urýchlili.