Jana Novotná & Vladan Glončák
Statistical NLP II
Assignemnt 3

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

# Task 1

Použili sme Brillov tagger z knižnice NLTK. Ako počiatočný tagger sme použili jednoduchý unigramový tagger, ktorý si zapamätá najčastejšie značku pre každé slovo trénovacích dat a pre neznáme slová vráti špecialnú značku "None". Použili sme počiatočnú množinu 24 vzorov[1] (template), ktoré pôvodne použil Brill. Pre angličtinu sme početpravidiel obmedzili[2] 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 spomaňuje.

## Výsledky

| **Angličtina** | **Přesnost** |
|---|---|
| 1. stupeň | 0.8910645842567726 |
| 2. stupeň | 0.88881069669247 |
| 3. stupeň | 0.8602301723689006 |
| 4. stupeň | 0.8931381232741104 |
| 5. stupeň | 0.8927185098345708 |
| Průměr | 0.885192417285 |
| $\sigma$ | 0.0125734981194 |
| Průmer triv. | 0.865374314735 |
| $\sigma$ triv. | 0.0143254860843 |

| **Čeština** | **Přesnost** |
|---|---|
| 1. stupeň | 0.7738987193793506 |
| 2. stupeň | 0.7888304373239811 |
| 3. stupeň | 0.7535799207397622 |
| 4. stupeň | 0.7922554742597533 |
| 5. stupeň | 0.7880596223212392 |
| Průměr | 0.779324834805 |
| $\sigma$ | 0.0124922171647 |
| Průmer triv. | 0.738888267001 |
| $\sigma$ triv. | 0.0134321134672 |

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

# Task 2

## Nesupervizovaný HMM — Baum-Welch

Pre túto časť sme sa rozhodli použiť iba bigramové a unigramové štatistiky, pretože ak by sme nechceli použiť použiť nuly "natvrdo" zabralo by to príšerne veľa času pre angličtinu, pre češtinu by to bolo jednoducho nerealizovateľné, časovo ani pamäťovo.

[1] NLTK má tieto vzory ako súčasť knižnice
[2] Limit sa pri týchto počtoch vždy naplní, čiže limit je rovnaký ako počet pravidiel.