

## Implementácia slovníka

V tejto etape projektu bol vytvorený frekvenčný slovník slovenského jazyka určený pre automatické dopĺňanie v klávesnici OpenBoard a v prostredí Android. Cieľom bolo pripraviť vstupné dáta v tvare kompatibilnom s LatinIME a overiť ich použitie pri generovaní návrhov slov.

Výstup bol pripravený na ďalšie testovanie v OpenBoard.

## Spracovanie textu a filtrácia

Textové dáta boli spracované pomocou vlastných Python skriptov. Počas spracovania boli aplikované nasledujúce filtre:

- 1 normalizácia slov na malé písmená
- 2 odstránenie neplatných a technických tokenov
- 3 spájanie variantov slova podľa registra
- 4 odstránenie veľmi zriedkavých výrazov

Tým sa výrazne znížilo množstvo šumu v korpuse a zlepšila sa kvalita slovníka.

## Generovanie Combined Dictionary

Zo spracovaných frekvencií bol vytvorený súbor vo formáte Combined Dictionary. Každé slovo dostalo normalizovanú frekvenciu a pre najčastejšie dvojice slov boli doplnené aj bigramy, ktoré zlepšujú kontextové návrhy pri písaní.

Počas implementácie bolo potrebné riešiť aj veľkosť výsledného slovníka.

Výsledný formát zostal kompatibilný s nástrojmi LatinIME.

## Testovanie a výsledok

Vygenerovaný slovník bol prevedený do binárneho formátu a integrovaný do OpenBoard. Testovanie prebiehalo v Android Emulatori s nastavenou slovenskou klávesnicou, pričom sa overilo správne načítanie slovníka aj návrhov slov. Po úpravách filtrácie a veľkosti dát sa podarilo dosiahnuť funkčné automatické dopĺňanie slovenských slov.