**FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY**

**UNIVERZITA KOMENSKÉHO**

**BRATISLAVA**

**ŠPECIFIKÁCIA POŽIADAVIEK NA SOFTVÉR**

Analýza lingvistických dát

zimný semester 2018/2019

Pavol Freivolt

Tamara Savková

Júlia Gablíková

Michal Knor

# Obsah

[Obsah 2](#_Toc527979452)

[1. Úvod 4](#_Toc527979453)

[1.1 Predmet špecifikácie 4](#_Toc527979454)

[1.2 Rozsah projektu 4](#_Toc527979455)

[1.3 Slovník pojmov 4](#_Toc527979456)

[1.4 Odkazy 4](#_Toc527979457)

[1.5 Prehľad nasledujúcich kapitol 5](#_Toc527979458)

[2. Všeobecný popis 5](#_Toc527979459)

[2.1 Perspektíva produktu 5](#_Toc527979460)

[2.2 Funkcie produktu 5](#_Toc527979461)

[2.3 Charakteristika používateľov 5](#_Toc527979462)

[2.4 Všeobecné obmedzenia 6](#_Toc527979463)

[2.5 Predpoklady a závislosti 6](#_Toc527979464)

[3. Špecifikácia požiadaviek 6](#_Toc527979465)

[3.1 Používateľské rozhranie 6](#_Toc527979466)

[3.1.1 Vkladanie textu priamo 6](#_Toc527979467)

[3.1.2 Vkladanie textu vo forme súboru 6](#_Toc527979468)

[3.1.3 Výber jazyka 6](#_Toc527979469)

[3.1.4 Výber kódovania 6](#_Toc527979470)

[3.1.5 Získanie výstupu 6](#_Toc527979471)

[3.2 Vstupný text 6](#_Toc527979472)

[3.3 Spracovanie textu 7](#_Toc527979473)

[3.3.1 Vyčistenie textu od nadbytočných znakov 7](#_Toc527979474)

[3.3.2 Zmena veľkosti písmen 7](#_Toc527979475)

[3.3.3 Rozdelenie textu na slabiky 7](#_Toc527979476)

[3.3.4 Zistenie početnosti slabík 7](#_Toc527979477)

[3.3.5 Zistenie dĺžky slabík 7](#_Toc527979478)

[3.3.6 Zistenie frekvencie výskytu dĺžok slabík vrátane opakovaných výskytov 7](#_Toc527979479)

[3.3.7 Zistenie frekvencie výskytu dĺžok slabík bez opakovaných výskytov 7](#_Toc527979480)

[3.4 Výstupné súbory 7](#_Toc527979481)

[3.4.1 Text rozdelený na slabiky 7](#_Toc527979482)

[3.4.2 Tabuľka s početnosťou a dĺžkou slabík 8](#_Toc527979483)

[3.4.3 Frekvenčná tabuľka dĺžok slabík berúc do úvahy všetky výskyty slabík 8](#_Toc527979484)

[3.4.4 Frekvenčná tabuľka dĺžok slabík berúc do úvahy výskyt slabík v texte iba raz 8](#_Toc527979485)

[3.4.5 Postupnosť dĺžok slabík 8](#_Toc527979486)

[3.5 Používateľská príručka 8](#_Toc527979487)

[3.5.1 Lokálne použitie konzolovej aplikácie 8](#_Toc527979488)

[3.5.2 Lokálne pridanie nového jazyka 8](#_Toc527979489)

[3.6 Správcovská príručka 8](#_Toc527979490)

[3.6.1 Inštalácia aplikácie na webový server 8](#_Toc527979491)

[3.6.2 Pridanie nového jazyka 8](#_Toc527979492)

[3.7 Ostatné požiadavky 9](#_Toc527979493)

[3.7.1 Jednoduchosť 9](#_Toc527979494)

[3.7.2 Voľná dostupnosť 9](#_Toc527979495)

# 1. Úvod

## 1.1 Predmet špecifikácie

V tejto špecifikácii požiadaviek sa popisujú používateľské, funkčné a parametrické požiadavky na aplikáciu, ktorá rozdeľuje slová na slabiky. Špecifikácia je určená pre zadávateľa a tím, ktorý bude aplikáciu implementovať. Zároveň je súčasťou dohody medzi objednávateľom a dodávateľom. Špecifikácia bude slúžiť ako východisko pre vyhodnocovanie správnosti softvéru.

## 1.2 Rozsah projektu

Výsledná aplikácia bude mať dve časti, ktoré sa budú dopĺňať - konzolovú aplikáciu a webové rozhranie. Konzolová aplikácia bude kompatibilná s operačnými systémami typu Microsoft Windows, Ubuntu a macOS. Používateľ bude môcť odoslať do aplikácie text v jednom z podporovaných slovanských jazykov. Aplikácia mu okrem iného vygeneruje ako výstup súbor, v ktorom je upravený vstupný text rozdelený na slabiky a ďalšie súbory, ktoré rozanalyzujú text vzhľadom na dĺžky a početnosti vyskytujúcich sa slabík. Úprava vstupného textu spočíva v tom, že v texte sú ponechané len samotné slová (neobsahuje napr. interpunkciu) a všetky písmená sú zmenené na malé. Aplikácia nenavracia ako výstup pôvodný text rozdelený na slabiky, vracia výlučne jeho upravenú verziu.

## 1.3 Slovník pojmov

1. Podporovaný slovanský jazyk – čeština, poľština, slovinčina, ruština, bieloruština, ukrajinčina, macedónčina, chorvátčina, bulharčina, horná a dolná lužická srbčina, srbčina
2. Konfiguračný súbor jazyka – súbor, ktorý popisuje jednotlivé jazyky a pravidlá, podľa ktorých sa v týchto jazykoch delia slová na slabiky, pričom formát tohto súboru je pevne daný a pripravený tak, aby s ním aplikácia vedela priamo pracovať

## 1.4 Odkazy

1. product\_specification\_2018\_10\_12.docx - súbor od zadávateľa so všeobecnými pravidlami delenia slov na slabiky
2. language\_year\_month\_day.docx - 12 súborov od zadávateľa so špecifikáciou jednotlivých jazykov (názvy súborov zodpovedajú uvedenej šablóne)

## 1.5 Prehľad nasledujúcich kapitol

Druhá kapitola dokumentu popisuje perspektívu a funkcionalitu produktu, charakterizuje používateľov a uvádza tiež obmedzenia kladené na produkt. Tretia kapitola podrobne popisuje jednotlivé požiadavky zadávateľa na vizuálnu i na funkcionálnu časť produktu, obsahuje detailnejší opis vstupných a výstupných súborov, priložených príručiek a rozoberá tiež proces spracovania súborov.

# 2. Všeobecný popis

## 2.1 Perspektíva produktu

Produktom je aplikácia slúžiaca ako nástroj pre analýzu lingvistických dát. Aplikácia dostane vstup od používateľa, spracuje ho a ponúkne výstup na stiahnutie. Produkt rozdeľuje vstupný text v jednom z podporovaných slovanských (neskôr prípadne aj ďalších) jazykov na slabiky podľa pravidiel jazyka tohto textu a spraví tiež analýzu výsledku. Vstup určený na spracovanie je možné vložiť do textového okna alebo nahrať ako textový súbor. Text vkladaný do textového okna musí byť v kódovaní UTF-8. Pri nahrávaní textového súboru treba vybrať jeho jazyk a kódovanie z ponuky. Po spracovaní vstupu aplikácia vráti niekoľko výstupov: textový súbor rozdelený na slabiky a excelové súbory s analýzami vstupu — tabuľka s početnosťou a dĺžkou slabík, frekvenčná tabuľka dĺžok slabík berúc do úvahy všetky výskyty slabík, frekvenčná tabuľka dĺžok slabík berúc do úvahy výskyt slabík v texte iba raz a postupnosť dĺžok slabík. Výstupné súbory sa dajú stiahnuť.

## 2.2 Funkcie produktu

Hlavnou funkciou produktu je rozdelenie slov textu na slabiky podľa pravidiel príslušného jazyka. Aplikácia pred spracovaním upraví vstup - vyčisti text od zbytočných znakov a zmení veľkosti písmen na malé. Následne rozdelí slová na slabiky, zistí početnosti i dĺžky slabík, frekvencie výskytov dĺžok slabík vrátane, aj bez opakovaných výskytov jednotlivých slabík a výsledky vráti vo forme výstupných súborov, ktoré si bude používateľ môcť stiahnuť.

## 2.3 Charakteristika používateľov

Stránka bude mat používateľov dvoch typov.

Návštevník – bude vedieť vkladať text alebo súbor, zvoliť vybrané vlastnosti textu a stiahnuť výstup po spracovaní. Aplikácia sa bude používať vedcami pre lingvistickú analýzu textov na základe rozdelenia slov na slabiky. Návštevníkovi postačujú začiatočnícke skúsenosti s prácou s počítačom.

Správca systému – udržuje systém v prevádzke, inštaluje na server a pridáva nové jazyky vrátane vytvorenia ich konfiguračných súborov. Časť správcovských úkonov je sprístupnená lokálne aj používateľovi, ktorý sa ich využívaním stáva správcom svojho systému.

## 2.4 Všeobecné obmedzenia

Aplikácia nevie rozpoznať jazyk ani kódovanie vstupu.

## 2.5 Predpoklady a závislosti

Aplikácia vráti správny výstup len v prípade správne zvoleného jazyka a kódovania vstupu, pretože spôsob spracovania textu závisí na týchto údajoch.

# 3. Špecifikácia požiadaviek

## 3.1 Používateľské rozhranie

Aplikácia bude používateľovi poskytovať možnosť vložiť vstupný text, vybrať jazyk a kódovanie vloženého textu a stiahnuť výstupné súbory.

### 3.1.1 Vkladanie textu priamo

Aplikácia bude obsahovať textové okno, do ktorého sa bude dať vložiť text určený na spracovanie.

### 3.1.2 Vkladanie textu vo forme súboru

Aplikácia bude ponúkať možnosť nahrať textový súbor obsahujúci text určený na spracovanie.

### 3.1.3 Výber jazyka

Pred spracovaním textu (nech už je vložený akokoľvek) si používateľ bude musieť vybrať jazyk textu, ktorý vložil. Pri výbere nesprávneho jazyka bude výstup chybný. Ponuka bude obsahovať zoznam podporovaných slovanských jazykov, pričom neskôr môžu byť pridané aj ďalšie podporované jazyky.

### 3.1.4 Výber kódovania

V prípade vloženia textu vo forme súboru si používateľ bude musieť vybrať jeho kódovanie. Pri zvolení nesprávneho kódovania nie je zaručená správnosť výstupu. V prípade vloženia textu do textového okna je nutné, aby text bol v kódovaní UTF-8.

### 3.1.5 Získanie výstupu

Aplikácia po spracovaní vstupu umožní používateľovi stiahnuť výstupné súbory jednotlivo alebo spolu.

## 3.2 Vstupný text

Hlavným vstupom aplikácie bude text (priamo vložený používateľom alebo získaný z textového súboru) ľubovoľnej dĺžky napísaný v cyrilike alebo latinke. Okrem znakov abecedy vybraného jazyka môže obsahovať aj iné znaky (napr. číslice, interpunkciu a podobne).

## 3.3 Spracovanie textu

Pred vyprodukovaním výstupu bude aplikácia musieť najprv upraviť vstupný text a následne upravený text spracovať rôznymi spôsobmi.

### 3.3.1 Vyčistenie textu od nadbytočných znakov

Aplikácia odstráni z textu všetky znaky, ktoré podľa súborov poskytnutých zadávateľom nepatria do abecedy vybraného jazyka. Okrem písmen abecedy zostanú iba medzery.

### 3.3.2 Zmena veľkosti písmen

Aplikácia zmení všetky veľké písmená na malé.

### 3.3.3 Rozdelenie textu na slabiky

Podľa pravidiel špecifických pre jazyk textu (zvolený používateľom) aplikácia rozdelí slová na slabiky. Pravidlá pre jednotlivé jazyky sú dané v súboroch poskytnutých zadávateľom a tie sú následne upravené do konfiguračných súborov pripravených pre spracovanie aplikáciou.

### 3.3.4 Zistenie početnosti slabík

Pre jednotlivé slabiky vyskytujúce sa v texte aplikácia zistí ich početnosť.

### 3.3.5 Zistenie dĺžky slabík

Pre jednotlivé slabiky vyskytujúce sa v texte aplikácia zistí ich dĺžku. Pravidlá na zisťovanie dĺžky slabiky v jednotlivých jazykoch sú definované v súboroch poskytnutých zadávateľom a konfiguračných súboroch z toho vyhotovených.

### 3.3.6 Zistenie frekvencie výskytu dĺžok slabík vrátane opakovaných výskytov

Aplikácia vypočíta počet výskytov dĺžok slabík, pričom zarátava aj opakované výskyty konkrétnej slabiky s danou dĺžkou. Príklad: „mama“ obsahuje dvakrát slabiku „ma“, a teda do počtu výskytov slabík s dĺžkou 2 prispieva dvoma výskytmi.

### 3.3.7 Zistenie frekvencie výskytu dĺžok slabík bez opakovaných výskytov

Aplikácia vypočíta počet výskytov dĺžok slabík, pričom sa nebudú zarátavať opakované výskyty konkrétnej slabiky s danou dĺžkou – teda každá slabika prispeje len raz. Príklad: „mama“ obsahuje dvakrát slabiku „ma“, a teda do počtu výskytov slabík s dĺžkou 2 prispieva iba jedným výskytom.

## 3.4 Výstupné súbory

Výsledky spracovania textu budú prezentované používateľovi formou výstupných súborov.

### 3.4.1 Text rozdelený na slabiky

Aplikácia vytvorí textový súbor, v ktorom bude vstupný text už po rozdelení na slabiky, pričom slabiky budú od seba oddelené pomlčkami.

### 3.4.2 Tabuľka s početnosťou a dĺžkou slabík

Aplikácia vytvorí excelový súbor, kde v prvom stĺpci budú slabiky, ktoré sa v texte vyskytli, v druhom stĺpci ich početnosti a v treťom ich dĺžky. Slabiky budú zoradené podľa početnosti zostupne.

### 3.4.3 Frekvenčná tabuľka dĺžok slabík berúc do úvahy všetky výskyty slabík

Aplikácia vytvorí excelový súbor, pričom v prvom stĺpci bude uvedené, aké dĺžky slabík sa vo vstupnom texte vyskytli (usporiadané vzostupne). Druhý stĺpec bude obsahovať počet výskytov týchto dĺžok (vrátane opakovaných výskytov).

### 3.4.4 Frekvenčná tabuľka dĺžok slabík berúc do úvahy výskyt slabík v texte iba raz

Aplikácia vytvorí excelový súbor, pričom v prvom stĺpci bude uvedené, aké dĺžky slabík sa vo vstupnom texte vyskytli (usporiadané vzostupne). Druhý stĺpec bude obsahovať počet výskytov týchto dĺžok (bez opakovaných výskytov).

### 3.4.5 Postupnosť dĺžok slabík

Aplikácia vytvorí textový súbor podobný textu rozdelenému na slabiky, avšak namiesto jednotlivých slabík budú v tomto súbore napísané ich dĺžky. Čísla budú oddelené pomlčkami. Príklad: „mama“ z pôvodného textu bude nahradené postupnosťou „2-2“.

## 3.5 Používateľská príručka

K aplikácii bude priložený textový súbor obsahujúci návody určené pre používateľa aplikácie.

### 3.5.1 Lokálne použitie konzolovej aplikácie

Príručka bude obsahovať podrobný návod, ako sfunkčniť konzolovú aplikáciu na počítači používateľa, ako ju spustiť a ďalej obsluhovať.

### 3.5.2 Lokálne pridanie nového jazyka

Používateľ bude môcť lokálne rozšíriť konzolovú aplikáciu o nový jazyk, ktorý bude následne zaradený medzi podporované jazyky. Návod bude okrem iného obsahovať aj popis štruktúry konfiguračného súboru jazyka, ktorý bude potreba pri pridávaní vytvoriť. Návod neobsahuje úpravu webového rozhrania.

## 3.6 Správcovská príručka

K aplikácii bude priložený textový súbor obsahujúci návody určené pre správcu aplikácie.

### 3.6.1 Inštalácia aplikácie na webový server

Príručka bude obsahovať podrobný návod, ako sfunkčniť aplikáciu na webovom serveri.

### 3.6.2 Pridanie nového jazyka

Správca bude môcť rozšíriť celú aplikáciu o nový jazyk, ktorý bude následne zaradený medzi podporované jazyky. Návod bude okrem iného obsahovať aj popis štruktúry konfiguračného súboru jazyka, ktorý bude potreba pri pridávaní vytvoriť. Súčasťou návodu je úprava webového rozhrania aplikácie.

## 3.7 Ostatné požiadavky

### 3.7.1 Jednoduchosť

Užívateľské prostredie bude prehľadné a ľahko ovládateľné. Aplikácia sa bude dať spustiť bez použitia iných aplikácií.

### 3.7.2 Voľná dostupnosť

Aplikácia bude vytvorená pomocou voľne dostupného legálneho softvéru.