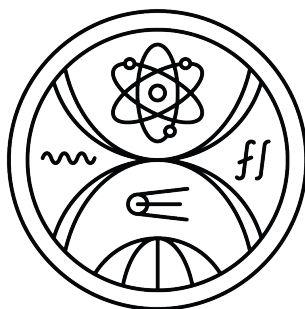


UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



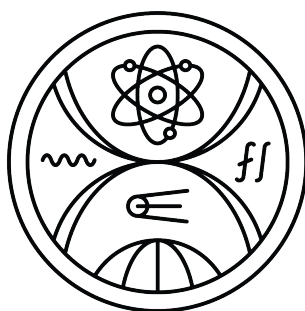
MODEL DISTRIBÚCIE INTERPUNKČNÝCH  
ZNAMIENOK V RÔZNYCH TEXTOCH  
DIPLOMOVÁ PRÁCA

2025

BC. MARTIN HOŠEK



UNIVERZITA KOMENSKÉHO V BRATISLAVE  
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



# MODEL DISTRIBÚCIE INTERPUNKČNÝCH ZNAMIENOK V RÔZNYCH TEXTOCH

DIPLOMOVÁ PRÁCA

Študijný program: Aplikovaná informatika  
Študijný odbor: Aplikovaná informatika  
Školiace pracovisko: Katedra aplikovanej informatiky  
Školiteľ: doc. RNDr. Mária Markošová, PhD.

Bratislava, 2025  
Bc. Martin hošek





Univerzita Komenského v Bratislave  
Fakulta matematiky, fyziky a informatiky

## ZADANIE ZÁVEREČNEJ PRÁCE

**Meno a priezvisko študenta:** Bc. Martin Hošek  
**Študijný program:** aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)  
**Študijný odbor:** informatika  
**Typ záverečnej práce:** diplomová  
**Jazyk záverečnej práce:** slovenský  
**Sekundárny jazyk:** anglický

**Názov:** Model distribúcie interpunkčných znamienok v rôznych textoch  
*Model of punctuation mark distributions in various texts*

**Anotácia:** Študent naprogramuje, alebo použije už naprogramovanú aplikáciu na hľadanie distribúcie interpunkčných znamienok v textoch v rôznych jazykoch. Ak študent použije hotovú aplikáciu, musí k nej doprogramovať v šetko potrebné na analýzu distribúcie a na tvorbu jej modelu. Študent zanalyzuje distribúcie z textov toho istého autora v rôznych časových obdobiach aby zistil, ako sa táto vlastnosť textu mení a nakoľko ostáva stabilná. Takisto zanalyzuje tieto distribúcie v textoch ľudí s afáziou. Navrhne matematický model tejto distribúcie.

**Cieľ:** = naprogramovať aplikáciu, alebo doprogramovať potrebné nástroje do aplikácie na analýzu distribúcií interpunkčných znamienok v textoch a vytvoriť matematický model opisujúci získané výsledky

**Literatúra:** Kulig and others. In narrative texts punctuation marks obey the same statistics as words, Information Sciences  
Volume 375, 1 January 2017, Pages 98-113

**Vedúci:** doc. RNDr. Mária Markošová, PhD.  
**Katedra:** FMFI.KAI - Katedra aplikovanej informatiky  
**Vedúci katedry:** doc. RNDr. Tatiana Jajcayová, PhD.  
**Dátum zadania:** 23.02.2024

**Dátum schválenia:** 21.11.2024  
prof. RNDr. Roman Ďurikovič, PhD.  
garant študijného programu

.....  
študent

.....  
vedúci práce

**Pod'akovanie:** Tu môžete poďakovať školiteľovi, prípadne ďalším osobám, ktoré vám s prácou nejako pomohli, poradili, poskytli dáta a podobne.

# Abstrakt

V tejto práci analyzujeme distribúciu interpunkčných znamienok v rôznych textoch. Naším cieľom je zistiť, aké sú rozdiely v používaní týchto znamienok v rôznych typoch textov a aké faktory môžu ovplyvniť ich výskyt. Analyzujeme texty ľudí s afáziou.

**Kľúčové slová:** afázia, zipov zákon, interpunkčné znamienka, distribúcia interpunkčných znamienok, frekvencia slov

# **Abstract**

In this work, we analyze the distribution of punctuation marks in various texts. Our goal is to determine the differences in the use of these marks in different types of texts and what factors may influence their occurrence. We analyze texts of people with aphasia.

**Keywords:**





# Obsah

<b>Úvod</b>	<b>1</b>
<b>1 Program - implementácia</b>	<b>3</b>
1.1 Načítanie PDF súboru a extrakcia textu po znakoch . . . . .	3
1.2 Analýza získaného textu a rozobratie textu na jednotlivé tokeny . . . .	3
1.3 Vyrábanie potrebných (grafových) štruktúr a ich interpretácia . . . . .	3
<b>2 Analýza textov</b>	<b>5</b>
2.1 Analýza frekvencie v texte EMR: Na Západe nič nové . . . . .	5



# Úvod

V tejto práci budeme skúmať distribúcie interpunkčných znamienok v rôznych textoch. Budeme sa zameriavať na analýzu textov ľudí s afáziou. Naším cieľom je zistiť, aké sú rozdiely v používaní týchto znamienok v rôznych typoch textov a aké faktory môžu ovplyvniť ich výskyt.

Budeme využívať literatúru [3] a [1].



# Kapitola 1

## Program - implementácia

V tejto kapitole popíšeme ako sme využili knižnicu [2] na analýzu a interpretáciu textu zo vstupného súboru formátu PDF.

- 1.1 Načítanie PDF súboru a extrakcia textu po znkoch
- 1.2 Analýza získaného textu a rozobratie textu na jednotlivé tokeny
- 1.3 Vyrábanie potrebných (grafových) štruktúr a ich interpretácia



# Kapitola 2

## Analýza textov

V tejto kapitole popíšeme zistenia ohľadom frekvencie interpunkčných znamienok vybranej literatúry..

### 2.1 Analýza frekvencie v texte EMR: Na Západe nič nové

Text románu Na západe nič nové (All Quiet on the Western Front [4]) od autora Ericha Maria Remarquua je napísaný v anglickom jazyku.





# Literatúra

- [1] Dynamika sietí. In Vladimír Kvasnička, Jozef Pospíchal, Jiří Navrátil, Bohumil Lacko, and Peter Trebatický, editors, *Umelá inteligencia a kognitívna veda II*, pages 321–377. STU, 2010.
- [2] Apache Software Foundation. Apache pdfbox library. <https://pdfbox.apache.org/download.html>, 2024. Version 3.0.4.
- [3] Kulig et al. In narrative texts punctuation marks obey the same statistics as words. *Information Sciences*, 375:98–113, Január 2017.
- [4] Erich Maria Remarque. *All Quiet on the Western Front*. Ballantine Books, 1929. Dostupné na: [https://www.glscott.org/uploads/2/1/3/3/21330938/aqwf-book\\_20size.pdf](https://www.glscott.org/uploads/2/1/3/3/21330938/aqwf-book_20size.pdf). [Cit.: 15/05/2025].