D08 – Database of Slovak and English legal terms (EN)

Project funded by the European Union Next GenerationEU through the Slovak Republic's Recovery and Resilience Plan under project no. 09I05-03-V02-00049.

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU
VAIA VÝSKUMNÁ A INOVAČNÁ AUTORITA
VÝSKUMNÁ AGENTÚRA

INTRODUCTION

The project "**Automation of Legal Text Analysis Based on Machine Learning**" (hereinafter referred to as "ALTAML") is a key initiative aimed at integrating innovative approaches in the field of data processing and subsequent analysis, with a focus on the field of law, which also includes information and communication technology law. The aim of this project is to develop and verify effective methods for the automated analysis of legal texts using machine learning techniques. This includes, in particular, the development of tools that facilitate the processing and analysis of large volumes of data in the form of various legal documents, as well as the extraction of relevant information (attributes) from such documents, including the identification of key terms, references to legislation and other attributes.

The aim of the ALTAML project is therefore to contribute to a more efficient approach to legal information and to speed up legal processes, thereby ensuring a higher degree of legal certainty and improving the accessibility of legal texts, the results of their analysis and relevant legal information for both professionals and the general public.

As part of the KPB2-4 work package, a database of Slovak and English legal terms (dataset) was developed. This document contains a description of the structure and methodology of creating the database of legal terms, as well as their definitions in Slovak and English in .csv, .xlsx and .json formats.

1. Dictionary of Slovak and English Legal Terms

The database is stored in .csv, .xlsx and .json formats. It contains legal terms collected from online sources such as **Slov-Lex** and **Najprávo.sk**, as well as from selected section headings in the Collection of Laws. These headings were then filtered using a combination of statistical and semantic metrics, as well as by prompting large language models.

The database also contains definitions of legal terms in Slovak, with some terms having definitions available from multiple sources, whether from online databases or directly from the texts of paragraphs in which the terms appear as headings. The database also includes English legal terms and their definitions, which were obtained by translation using the large language model **LLaMA 3.3**.

2. Features

The databases pravne_pojmy_sk_en.csv and pravne_pojmy_sk_en.xlsx contain only lists of terms, where the **SK** column represents Slovak legal terms and the **EN** column their English equivalents. These files do not contain definitions or other metadata.

The legal_terms_sk_en.json database contains extended information on terms and includes the following fields:

- **_id**: the legal term itself in Slovak (e.g. 'premlčacia doba')

- **definitions**: a list of definitions corresponding to the term, each definition containing:
    - **source**: the source of the definition (e.g. 'najpravo')
    - **definition**: the definition itself in Slovak (e.g. "Premlčacia doba je ...")

If the value of the **source** field is "zz" (Collection of Laws) and the term was taken from the heading of a section, the **definition** field does not contain simple text, but has a more structured character:

- **law**: identification of the law (e.g. "61/1952")

- **title_tag_id**: identifier of the heading in the structure of the law (e.g. "regulation.heading-10.title")

- **definition**: definition of the term extracted from the text of the paragraph in Slovak

- **english**: information about the English equivalent of the term obtained using a large language model:
    - **term**: English equivalent of the legal term (e.g. "statute of limitations")
    - **definition**: English definition of the term created by translating and summarising Slovak definitions

3. Requirements

For basic use of the database, a tool for reading and editing files in .csv and .xlsx formats, such as text editors or spreadsheet processors (e.g. Microsoft Excel), is sufficient.

To work with an extended database containing definitions in .json format, it is advisable to use software designed for working with databases or structured data, such as **MongoDB**, **CouchDB**, **PostgreSQL**, or even advanced text editors.

4. Sources:

The sources of legal terms are publicly available online repositories, in particular **Slov-Lex** and **Najprávo.sk**, which also provide definitions for the terms. This set of terms was subsequently expanded to include selected terms obtained from the headings of sections of laws in the Collection of Laws.
This step was motivated both by the insufficient coverage of legal terminology in available online sources and by the observation that many sections are of a definitional nature. Since not all section headings are suitable as legal terms, they were filtered based on an analysis of their wording, the content of the section itself, and their general legal nature, including the use of large language models.
In cases where the text of a section was identified as definitional, the definition was extracted directly and assigned to the relevant term.
Due to the lack of high-quality sources of legal terms in English, the English equivalents were obtained by translation using a large language model. The relevant English definitions were created by translating and summarising all available Slovak definitions.
Although these definitions may not always be completely accurate in terms of terminology, as they are generated data, their primary use is in the creation of

an embedding vector representations of terms, where context and thematic consistency are more important than absolute linguistic accuracy.