PONK: Psaní orientované na klienta

Smlouva č. TQ01000526

Odborná zpráva o řešení projektu za rok 2023

Příloha k Průběžné zprávě za rok 2023

Předkládá hlavní řešitel projektu

Barbora Vidová Hladká

za účastníky

Univerzita Karlova - Matematicko-fyzikální fakulta

Frank Bold Society

MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

Frank Bold Society

Úvod

Cílem projektu je vytvořit nástroj pro automatické hodnocení srozumitelnosti českých právních textů. Hodnocení bude obsahovat kvantitativní posouzení srozumitelnosti a argumentační struktury textů a identifikaci úseků textů, které s ohledem na větší srozumitelnost vyžadují další autorovu pozornost. Cílovou skupinou nástroje PONK, Psaní Orientované Na Klienta, jsou autoři právních textů. K měření srozumitelnosti a k vyhledávání její jazykové realizace použijeme metody počítačového zpracování přirozeného jazyka a strojového učení.

Naplňování cílů projektu

Tématicky je projekt rozdělen do tří okruhů: Data, Anotace a Aplikace.

Okruh Data

Datová základna je klíčovou komponentou projektu, protože navrhujeme datově orientované řešení. Správa dat a její exploratorní analýza jsou obsahem okruhu Data. V prvním roce řešení jsme publikovali první verzi korpusu KUK, KUK 0.0 = výsledek TQ01000526-V1, který obsahuje české právní a administrativní texty v objemu 19 196 037 slov v 6 051 dokumentech. Tato kolekce je pilotní kolekcí pro analýzu srozumitelnosti úřední dokumentace a při jejím sestavování jsme vycházeli z materiálu, který byl součástí návrhu projektu a který obsahoval přehled vytipovaných zdrojů.

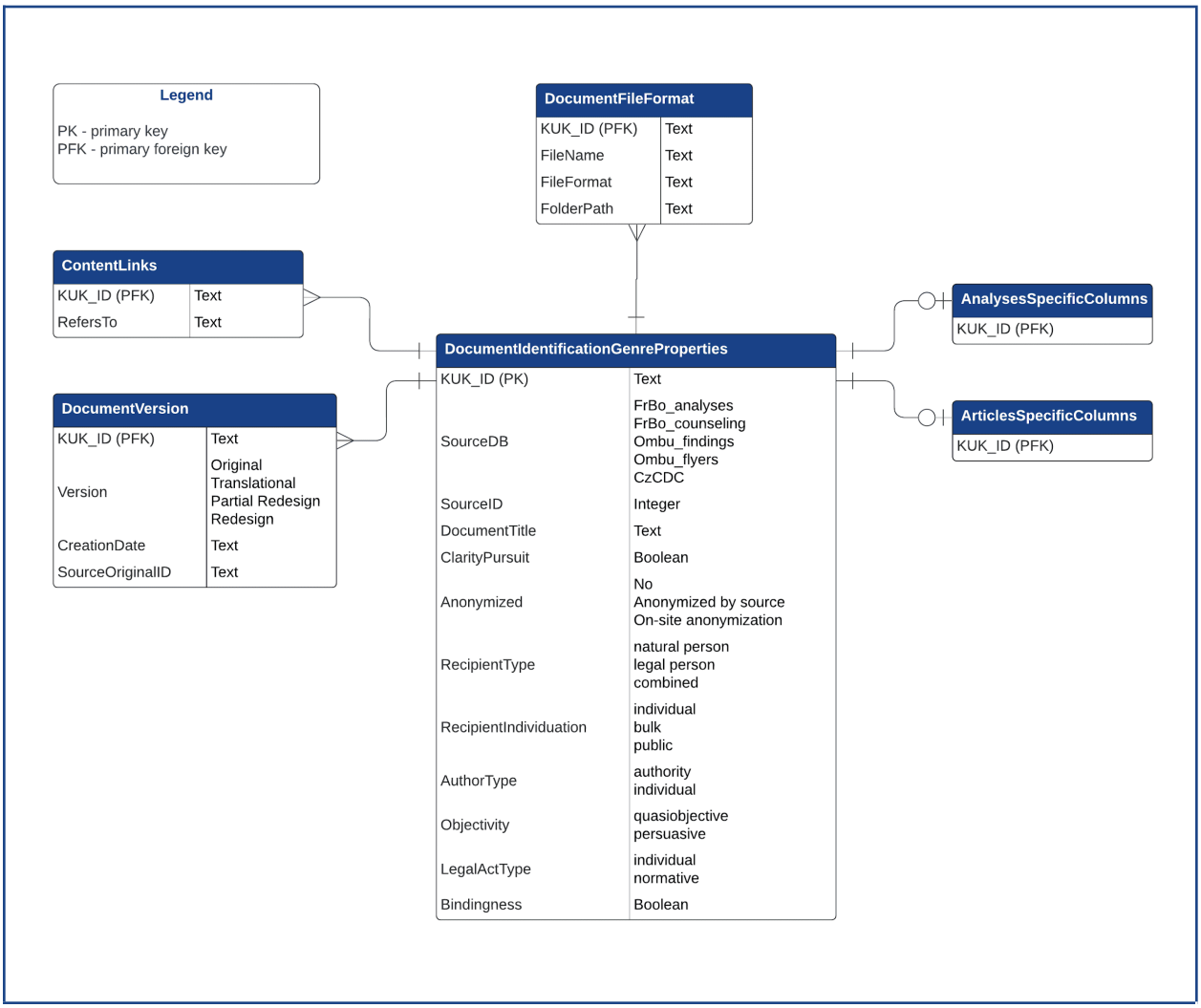
KUK 0.0

Do korpusu KUK 0.0 jsme zahrnuli (1) právní analýzy a zodpovězené dotazy z databáze interního aplikační garanta, (2) stanoviska a (3) letáky vydané Kanceláří veřejného ochránce práv, (4) soudní rozhodnutí tří nejvyšších českých soudů a (5) právní texty v několika parafrázích ze studie srozumitelnosti českých administrativních textů. Pro každou datovou sadu jsme zdokumentovali námi navrženou sadu metadat (viz Obrázek 1) a její formáty.[1] Mezi metatadaty je i binární atribut ClarityPursuit: některé dokumenty začleněné do KUK 0.0 byly napsány s cílenou snahou o jejich srozumitelnost (např. LiFR-Law), a to od zdroje, který důsledně uplatňuje politiku srozumitelných textů. Tyto dokumenty mají hodnotu atributu ClarityPursuit TRUE (PRAVDA), ostatní FALSE (NEPRAVDA). Míry srozumitelnosti budeme analyzovat a navrhovat v následujícícm roce řešení, a proto jsou stávající hodnoty atributu zcela předběžné.

Korpus KUK 0.0 je publikován v repozitáři LINDAT/CLARIAH-CZ, který plně respektuje pravidla FAIR a podporuje principy otevřené vědy (Open Science) pro přístup k vědeckým datům a

[1] Odkaz na dokumentaci korpusu https://ufal.mff.cuni.cz/grants/ponk/kuk.

výstupům vytvořených s veřejnou podporou.[2] Vlastnosti repozitáře jsou popsané v dokumentu Plán správy dat, který je součásti průběžné zprávy.



Obrázek 1: Konceptuální schéma metadat korpusu KUK 0.0

Okruh Anotace

Aplikace PONK bude využívat metody strojového učení, pro které bude potřeba vytvořit trénovací data. Pro naši úlohu to znamená ručně anotovat problematické jazykové jevy a argumentační strukturu textů vybraných z korpusu KUK. K anotaci jsme vybrali nástroj Gloss. Zvažovali jsme použití nástroje Brat, se kterým má řešitelský kolektiv z MFF UK zkušenosti z předchozích projektů, ale vlastnosti nástroje Gloss cílené na anotaci právních textů budou pro naše anotace vhodnější.

[2] Odkaz ke stažení korpusu http://hdl.handle.net/11234/1-5363.

Gloss

Gloss je anotační nástroj vyvinutý pro sémantickou anotaci zejména právních textů (viz Obrázek 2). Gloss byl vytvořen na Univerzitě v Pittsburghu a byl použit v několika studiích a výzkumných projektech (Poudyal a kol., 2023), (Šavelka a kol., 2023). Gloss lze modifikovat na všech úrovních anotačního procesu, tj. od tvorby korpusu, přes definici anotačního schématu a typologie anotací, samotné anotace dokumentů uživateli až po kontrolu kvality anotací. Z tohoto důvodu systém umožňuje přidělení jednotlivých identit a různých rolí s různými pravomocemi zasahovat do samotné metodologie anotací, kde tyto role mohou být i vícenásobné (např. jeden uživatel může být jak anotátorem, tak editorem).

Anotace jsou v nástroji Gloss identifikovány rozsahem, tedy číslem počátečního a koncového znaku textu. Díky této metodě je možné anotovat i přesahující text, tedy konkrétní část textu je možné zahrnout do více anotačních typů. Anotačním typům je dále také možné přiřazovat atributy.

Gloss je vysoce modifikovatelný nástroj, kde každá z jednotlivých fází anotačního procesu může být z role editora téměř libovolně nastavena. Z uživatelského pohledu je výhodou vysoká uživatelská přívětivost i pro anotaci dlouhých částí textu (nástroj obsahuje v pravém panelu vizuální přehled dokumentu), či právě při tvorbě překrývajících se anotací. Anotátoři si mohou také libovolně vybírat z přednastaveného přiřazeného seznamu dokumentů k anotaci a vracet se k nim. Systém je také pravidelně zálohován.

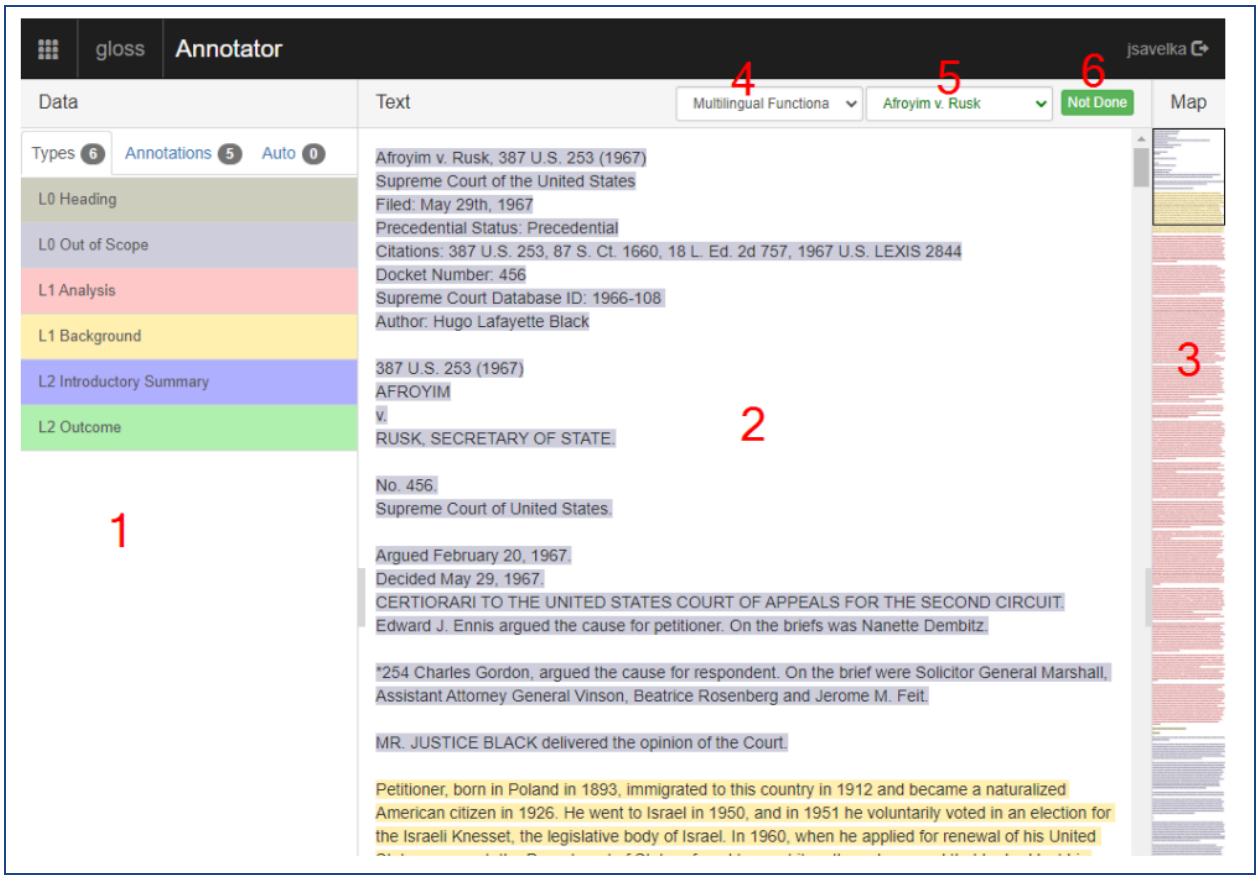
Brat

Brat je webový open-source nástroj pro anotování textu.[3] Umožňuje označit a klasifikovat libovolně dlouhé a rovněž překrývající se úseky textu a mezi těmito úseky anotovat vztahy různých druhů (Obrázek 3 ukazuje příklad anotace větné struktury).

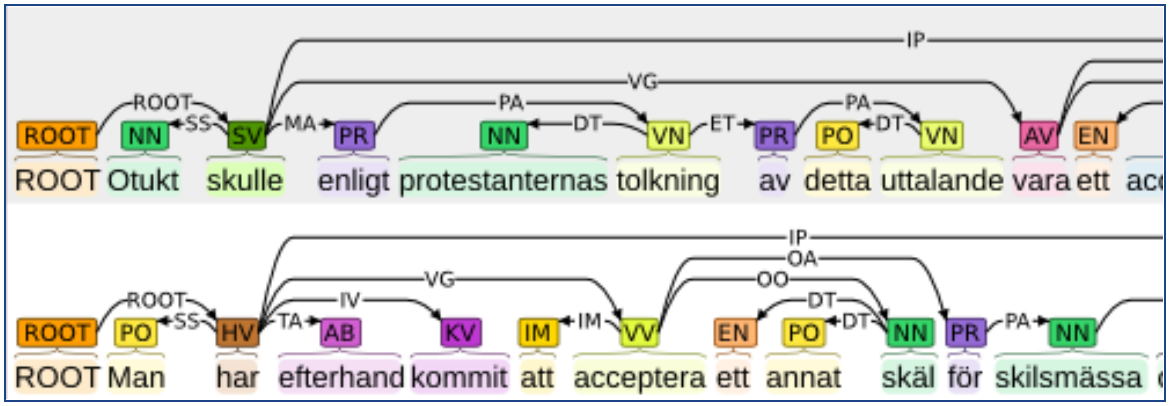
K výhodám Bratu patří intuitivní anotační prostředí, možnost lokální instalace serveru, široká přizpůsobitelnost potřebám konkrétní úlohy a velmi snadná konfigurace jak struktury a vzhledu anotace, tak i přístupu uživatelů. Pracoviště hlavního řešitele má s tímto nástrojem široké (a pozitivní) zkušenosti, kdy byl použit k anotaci větných subjektů, či v jiné úloze k anotaci citačních signálů v textu, citačních zdrojů a jejich propojení.

Způsob zobrazení označeného textu v Bratu je ovšem vhodný spíše pro krátké úseky, které nepřesahují příliš často hranici řádku. Zobrazení označených delších úseků textu je nešikovné (Obrázek 4) a z tohoto důvodu je Brat pro účely aktuálního projektu nevhodný.

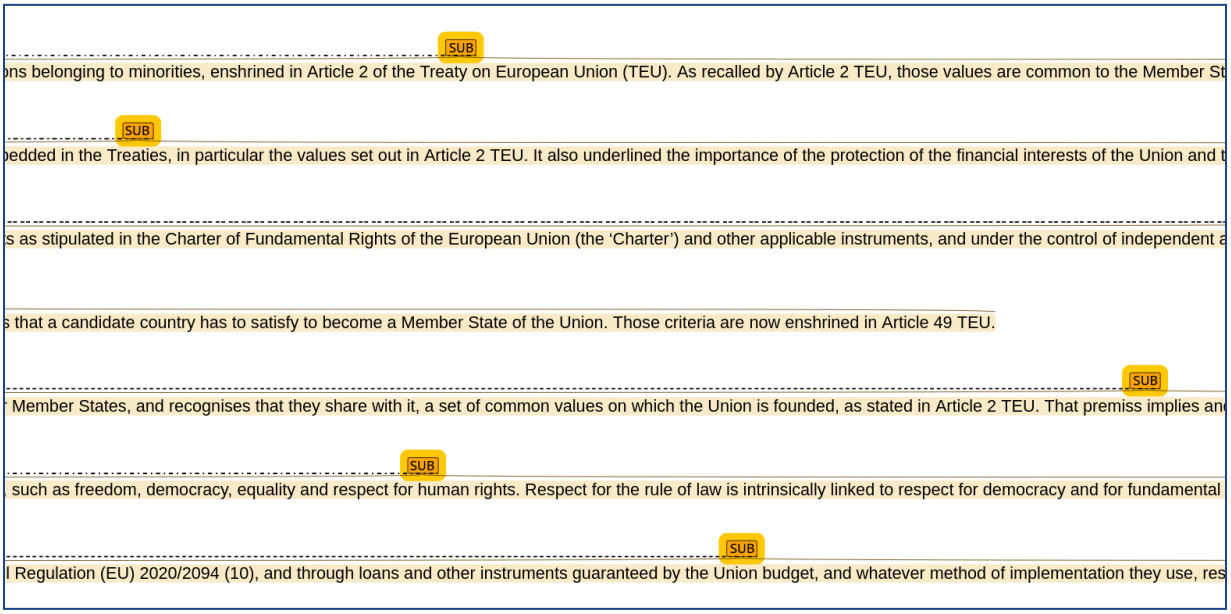
[3] https://brat.nlplab.org/index.html



Obrázek 2: Uživatelské rozhraní nástroje Gloss a jednotlivé součásti: 1. Typy anotací, 2. Textové okno, 3. Přehled dokumentu, 4. Výběr anotační úlohy, 5. Výběr dokumentu k anotaci, 6. Tlačítko k označení hotového dokumentu.



Obrázek 3: Příklad anotace větné struktury v Bratu.



Obrázek 4: Označení víceřádkového úseku textu v Bratu.

Okruh Aplikace

V okruhu Aplikace vznikne hlavní výsledek projektu, a sice nástroj PONK pro automatické hodnocení srozumitelnosti českých právních textů. Technicky bude implementován jako webová služba s rozhraním REST API.

Z diskusí o textech pro korpus KUK i o vstupech aplikace PONK bylo zřejmé, že bude nutné zajistit propojení nástroje PONK s anonymizační procedurou (anonymizérem). Ministerstvo spravedlnosti takový nástroj v minulosti vytvářelo, nicméně žádný není v tuto chvíli veřejně přístupný. Aktuálně probíhá vývoj nového nástroje není však známo kdy a v jaké podobě by byl k dispozici. Proto jsme navrhli a implementovali anonymizační nástroj MasKIT (TQ01000526-V9) jako webovou službu s rozhraním REST API, jehož využití není omezeno pouze pro nástroj PONK, ale je k dispozici pro libovolnou úlohu zpracování textů, která obsahují citlivá data podléhající ochraně. Technicky je ošetřeno, že se zpracovávané texty nikde neukládají.

Nástroj MasKIT vznikl jako reakce na potřebu od samého začátku projektu pracovat s texty obsahujícími citlivá data, přestože tento výsledek nebyl v návrhu projektu naplánován.

MasKIT

MasKIT je webový nástroj[4] a služba REST API pro anonymizaci českých textů se zaměřením na právní texty. MasKIT přijímá na vstupu český text (např. úřední rozhodnutí), v něm nalezne osobní informace (jména, adresy, rodná čísla, názvy firem, IČO, čísla pozemků atd.) a nahradí jejich výskyty náhodně zvolenými ekvivalenty.

Ke své činnosti využívá MasKIT dvou externích nástrojů vyvinutých Ústavem formální a aplikované lingvistiky MFF UK, a to UDPipe[5] a NameTag.[6] UDPipe analyzuje vstupní text a provede jeho tokenizaci (segmentace na slova), segmentaci na věty, lemmatizaci (generování základních tvarů slov), morfologické značkování a konečně syntaktickou analýzu vět ve formalizmu Universal Dependencies.[7] NameTag následně v textu vyhledá a klasifikuje jmenné entity (jména osob, adresy, názvy firem, geografická jména a mnoho dalších).

V takto zpracovaném textu MasKIT vyhledává osobní informace a nahrazuje je ekvivalenty z předem připraveného seznamu. Např. pro nahrazování křestních jmen má nástroj k dispozici 10 nejčastějších českých křestních jmen, obdobně pro další kategorie (jména ulic, názvy obcí, …).

---

[4] https://quest.ms.mff.cuni.cz/maskit/

[5] https://lindat.mff.cuni.cz/services/udpipe/

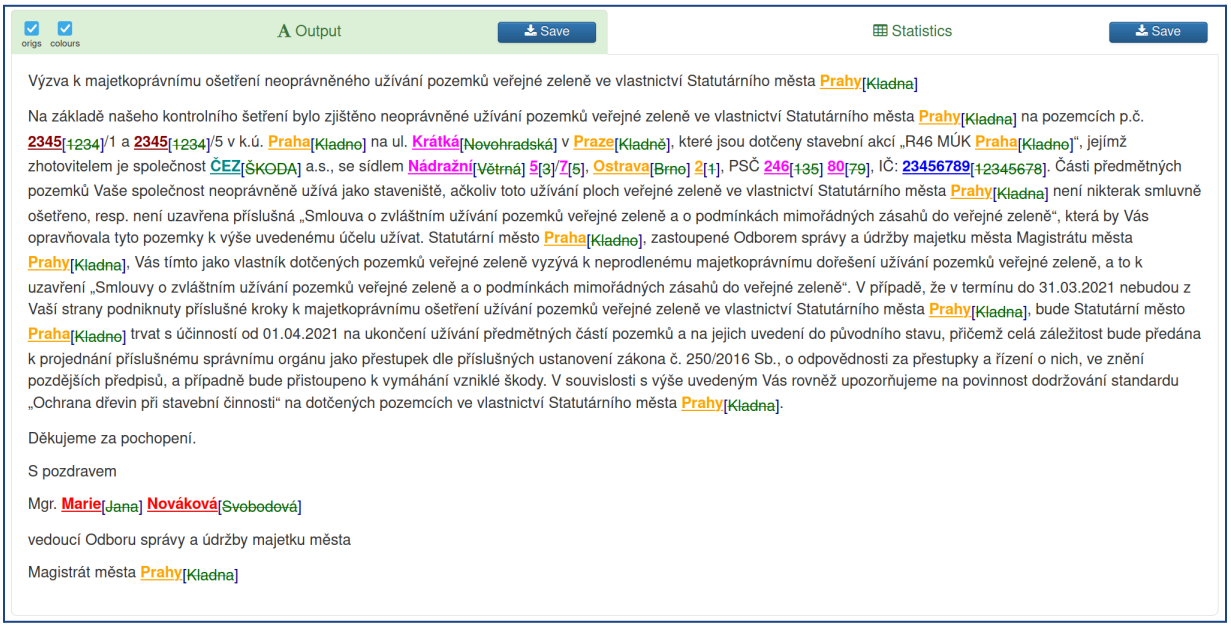
[6] https://lindat.mff.cuni.cz/services/nametag/

[7] https://universaldependencies.org/

Směrovací čísla, rodná čísla a další číselné identifikátory jsou nahrazovány jednoduchými variantami (např. směrovací číslo 123 45). Mužská a ženská příjmení jsou zpracována společně (tedy pokud systém nahradí např. příjmení Svoboda příjmením Novák, ve stejném dokumentu nahradí příjmení Svobodová příjmením Nováková).

Text s nahrazenými osobními informacemi je předložen uživateli ke kontrole a dalšímu zpracování (Obrázek 5). Uživatel má možnost zobrazit si text i s původními informacemi (jako je na Obrázku 5) či pouze s novými, s barevným zvýrazněním či bez zvýraznění.

MasKIT i externě volané služby zachovávají anonymitu, zpracovaný text tedy není nikde uložen.



Obrázek 5: Výsledek anonymizace nástrojem MasKIT[8]

Způsob zapojení členů konsorcia

V prvním roce řešení projektu, tj. v období září-prosinec 2023 se týmu podařilo splnit plánovaný výsledek TQ01000526-V1 a vytvořit jeden výsledek TQ01000526-V9 nad rámec plánu pro první rok řešení, a to díky pravidelným schůzkám a intenzivní spolupráci. Několika schůzek se účastnili i Barbora Kubíková a Jana Šamánková, se kterými jsme konzultovali datové sady z

[8] V tomto případě i původní text byl již anonymizován.

agendy Kanceláře veřejného ochránce práv. Obě budou v roce 2024 zapojeny do anotace textů z korpusu KUK. Schůzek se dále účastnili dva studenti MFF a FF UK, kteří přemýšlejí o tématech svých bakalářských prací navázaných na téma projektu PONK.

Všichni členové řešitelského týmu participovali na přípravě průběžné zprávy za rok 2023. Všichni členové řešitelského týmu jsou zapojeni do rešerše prací a výsledků tématicky blízkých projektu PONK. K jejich evidenci používáme citační software Zotero.[9] Dílčí odpovědnosti členů řešitelského týmu byly následující:

Univerzita Karlova - Matematicko-fyzikální fakulta

Barbora Vidová Hladká (klíčová osoba projektu) byla zodpovědná za koordinaci řešení projektu (včetně přípravy průběžné zprávy za rok 2023) a podílela se na publikaci korpusu KUK 0.0 (výsledek TQ01000526-V1) a na návrhu nástroje MasKIT (TQ01000526-V9). Dále byla odpovědná za sestavení Plánu správy dat.

Silvie Cinková (klíčová osoba projektu) koordinovala práci na korpusu KUK a navrhla a zdokumentovala jeho metadatové schéma. Podílela se na publikaci korpusu KUK 0.0 (výsledek TQ01000526-V1).

Jiří Mírovský (klíčová osoba projektu) navrhl a implementoval nástroj MasKIT (výsledek TQ01000526-V9) a připravil korpus KUK 0.0 (výsledekTQ01000526-V1) k publikaci v repozitáři LINDAT/CLARIAH-CZ.

Tereza Novotná (postdoc) připravila metadata sady CzCDC začleněné do korpusu KUK 0.0 (výsledek TQ01000526-V1) a vypracovala rešerši nástroje Gloss s ohledem na potřeby projektu.

Frank Bold Society

Michal Kuk (klíčová osoba projektu) připravil data a metadata sady FrBo začleněné do korpusu KUK 0.0 (výsledek TQ01000526-V1).

Kristýna Nguyen Zahálková (právnička právní poradny) organizovala a revidovala zdrojové dokumenty pro data sady FrBo začleněné do korpusu KUK 0.0 (výsledek TQ01000526-V1)

Lucie Petrová (stážistka právní poradny) poskytovala podporu při přípravě podkladů a práci členů týmů na korpusu KUK 0.0 (výsledek TQ01000526-V1)

[9] https://www.zotero.org/groups/5229972/ponk/library

Naplňování programu SIGMA z hlediska jeho zaměření

Zaměření projektu PONK spadá do hlavního tématu Odolnost společnosti a podtématu Odolnost společnosti vůči dezinformacím, mechanismy zvyšování důvěry občanů v demokratickou společnost veřejné soutěže TAČR programu SIGMA (DC3)

Dezinformace rozšiřují mezi lidmi nedůvěru v zavedené zdroje informací a veřejnou moc ve státě, což představuje pro demokratickou společnost hrozbu. V takové situaci je třeba intenzivně posilovat právní vědomí občanů. I když se názory na formování právního vědomí liší, shodují se, že právní vědomí zahrnuje znalosti o platnosti a oprávněnosti práva.

Úřední komunikace je stěžejní součástí komunikace státu vůči občanům. Zpravidla je při ní rozhodováno o jejich právech a povinnostech. Pro transparentnost procesů, zvýšení důvěry ve stát a právní stát je nezbytné, aby tato komunikace byl lidem dobře srozumitelná.

V projektu přistupujeme ke zvýšení srozumitelnosti práva prakticky tak, že na robustní kolekci právních dokumentů ohodnotíme jejich srozumitelnost ručně i automaticky, čímž vznikne unikátní vzdělávací materiál, a implementujeme webovou aplikaci dostupnou komukoli pro ověření srozumitelnosti libovolného právního dokumentu.

V prvním roce řešení projektu vznikla pilotní verze unikátní kolekce právních textů zaměřená na evaluaci srozumitelnosti, korpus KUK 0.0. Korpus je k dispozici ke stažení pod volnou licencí Creative Commons z repozitáře LINDAT/CLARIAH-CZ. Vznikl tak materiál k ruční analýze srozumitelnosti a k trénování modelů strojového učení. V této fázi projektu slouží jako vzdělávací materiál pro experty na srozumitelné psaní a pro experty v oblasti strojového učení. Široké veřejnosti bude předložen v uživatelsky přívětivějším prostředí v dalším roce řešení a následně zprostředkovaně v aplikaci automatického hodnocení srozumitelnosti PONK.

Literatura

- POUDYAL, Prakash, Jaromír ŠAVELKA, Aagje IEVEN, Marie Francine MOENS, Teresa GONCALVES a Paulo QUARESMA, 2020. ECHR: Legal Corpus for Argument Mining. In: *ArgMining 2020*: *Proceedings of the 7th Workshop on Argument Mining* [online]. Online: Association for Computational Linguistics, s. 67–75 [vid. 2023-10-19]. Dostupné z: https://aclanthology.org/2020.argmining-1.8.
- ŠAVELKA, Jaromí, Hannes WESTERMANN, Karim BENYEKHLEF, Charlotte S. ALEXANDER, Jayla C. GRANT, David Restrepo AMARILES, Rajaa El HAMDANI, Sébastien MEEÙS, Aurore TROUSSEL, Michał ARASZKIEWICZ, Kevin D. ASHLEY, Alexandra ASHLEY, Karl BRANTING, Mattia FALDUTI, Matthias GRABMAIR, Jakub HARAŠTA, Tereza NOVOTNÁ, Elizabeth TIPPETT a Shiwanni JOHNSON, 2021. Lex Rosetta: transfer of predictive

models across languages, jurisdictions, and legal domains. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* [online]. New York, NY, USA: Association for Computing Machinery, s. 129–138 [vid. 2023-10-19]. ICAIL '21. ISBN 978-1-4503-8526-8. Dostupné z: doi:10.1145/3462757.3466149.