PONK: Psaní orientované na klienta

Smlouva č. TQ01000526

Odborná zpráva o řešení projektu za rok 2024

Příloha k Průběžné zprávě za rok 2024

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

1. Naplňování cílů projektu

Srozumitelnost jazyka používaného ke komunikaci práva je důležitou součástí právního státu (Curtotti, McCreath, 2013). Nesrozumitelná právní pravidla vytvářejí nedůvěru v právní řád a státní instituce a znesnadňují adresátům práva jeho dodržování. Otázka srozumitelnosti práva úzce souvisí s otázkou přístupu k právu a spravedlnosti, což jsou ústavní principy (Serota, 2011). Kromě tohoto ústavněprávního aspektu má srozumitelnost právního textu širší pozitivní sociální důsledky - umožňuje adresátům zvládnout základní každodenní právní záležitosti, aniž by museli oslovovat právníka, což může být velmi nákladné. Zejména s rozšířením internetu, a tím i dostupnosti právních informací online, vyvstala otázka jejich srozumitelnosti jako druhý krok v liberalizaci právních služeb (Curtotti a kol. 2015). Přístup k relevantním a aktuálním právním informacím je předpokladem pro jejich pochopení, a tedy i k jednání na jejich základě.

Často jsou úřední nebo právní texty doručeny bez právní asistence, což ponechává jednotlivce - někdy se sníženou kognitivní schopností - bez pomoci a zároveň zodpovědné za porozumění svých práv a povinností. Mezi běžné scénáře patří pokuty, oznámení o poplatcích nebo přístup k sociálním dávkám. Veřejný ochránce práv v takových situacích sice nabízí bezplatnou asistenční službu, nicméně poptávka výrazně převyšuje jeho kapacitu, jak dokládá výroční zpráva za rok 2023.[1] Přitom srozumitelnost běžného úředního sdělení může výrazně snížit počet žádostí směřovaných na Veřejného ochránce práv, ale zejména významně zvýšit informovanost veřejnosti a důvěru v orgány veřejné moci, založenou na tom, že adresát právního dokumentu je schopen ho pochopit a vyřešit životní situaci samostatně a legislativně správně.

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 tak, aby vzniklo flexibilní řešení, které by zároveň podpořilo jeho přijetí právníky.

V nástroji PONK uplatňujeme datově orientovaný přístup k řešení, v němž jsme zvolili tři aspekty srozumitelnosti textu, a sice lexikální překvapení, stylistická vhodnost formulací a argumentační struktura. Tyto aspekty zároveň odpovídají dílčím modulům softwarového řešení nástroje PONK. Pokud je nám známo, takto pojatý přístup dosud nebyl v národním ani mezinárodním kontextu uplatněn. Architektuře řešení odpovídá i členění projektu na okruhy Data, Anotace a Aplikace.

[1] https://www.ochrance.cz/dokument/zpravy_pro_poslaneckou_snemovnu_2023/vyrocni-zprava-2023.pdf

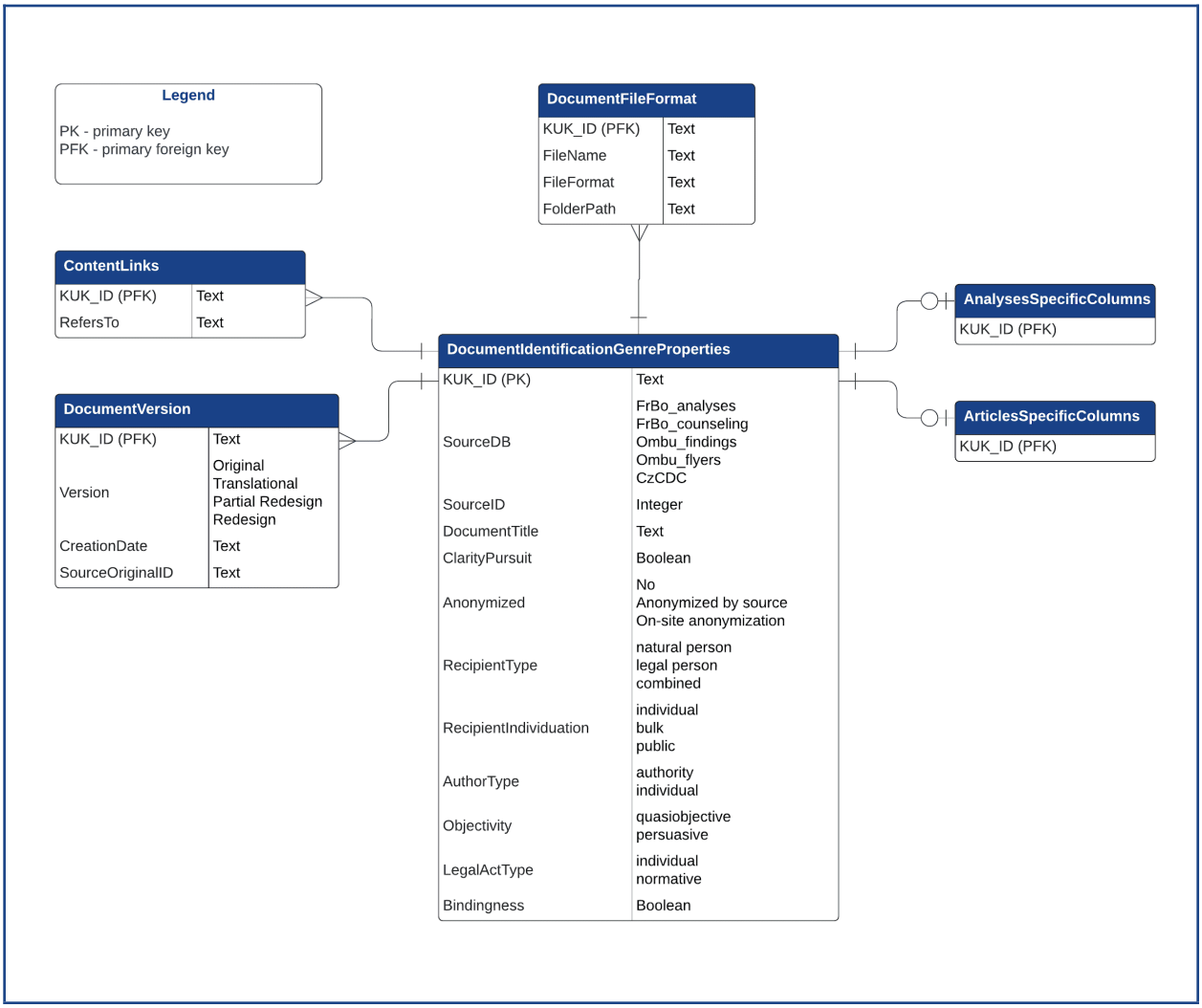
Ve druhém roce řešení projektu, leden-prosinec 2024, řešitelský tým splnil všechny tři naplánované výsledky

- KUKY 1.0 (SB - Specializovaná veřejná databáze, výsledek TQ01000526-V7)
- KUK 1.0 (SB - Specializovaná veřejná databáze, výsledek TQ01000526-V2)
- PONK 1.0 (Gfunk - Funkční vzorek, výsledek TQ01000526-V5)

Webová stránka projektu https://ufal.mff.cuni.cz/grants/ponk je rozcestníkem ke všem výsledkům.

1.1 Okruh Data

V rámci okruhu Data jsme v roce 2024 publikovali druhou verzi korpusu KUK, KUK 1.0 (výsledek TQ01000526-V2), a korpus KUKY 1.0 (výsledek TQ01000526-V7). Oba korpusy jsou publikovány v repozitáři LINDAT/CLARIAH-CZ, který plně respektuje pravidla FAIR a podporuje principy



Obrázek 1: Konceptuální schéma metadat korpusu KUK

otevřené vědy (Open Science) pro přístup k vědeckým datům a výstupům vytvořeným s veřejnou podporou.[2] Vlastnosti repozitáře jsou popsané v dokumentu Plán správy dat, který je součásti průběžné zprávy.

KUKY 1.0

Korpus KUKY 1.0 obsahuje 224 anotovaných textů částečně z veřejně dostupných zdrojů, částečně z anonymizovaných a redakčně upravených různých verzí přepracovávaných textů z kolekce školících materiálů Jany Šamánkové a Barbory Kubíkové. V metadatech korpusu (viz níže) zvlášť uvádíme vztahy mezi různými verzemi téhož dokumentu. Pozdější verze je vždy považována za srozumitelnější než dřívější verze, avšak je třeba zdůraznit, že ani výchozí verze nemusela být nutně nesrozumitelná.[3] V textech je ručně anotovaná argumentační struktura - podrobněji viz část Okruh Anotace.

KUK 1.0

Korpus KUK 1.0 je automaticky anotovaný nástrojem UDPipe pro morfologickou a syntaktickou analýzu textů a nástrojem NameTag pro detekci a klasifikaci pojmenovaných entit.[4] Konceptuální schéma metadat navržené pro KUK 0.0 nebylo nutné pro KUK 1.0 upravovat (viz obrázek č. 1). Obsahuje české právní a administrativní texty z první verze KUK 0.0 a z KUKY 1.0:

- data ze čtyř zdrojů spolu s metadaty, 6 275 dokumentů s 19 559 553 slovy
  - veřejné materiály interního aplikačního garanta (FrBo)
  - stanoviska veřejného ochránce práv (ESO)
  - informační letáky vydané kanceláří veřejného ochránce práv (OmbuFlyers)
  - texty z KUKY 1.0
- metadata pro data ze dvou externích korpusů, které byly publikovány v repozitáři LINDAT/CLARIAH-CZ před zahájením projektu PONK
  - Czech Court Decisions Corpus (CzCDC 1.0)[5]
  - Corpus of Paraphrased Czech Administrative Texts with Reading Comprehension for Readability Studies (LiFR-Law)[6]

1.2 Okruh Anotace

V rámci okruhu Anotace byla v roce 2024 ručně anotovaná argumentační struktura 224 úředních dokumentů začleněných do korpusu KUKY 1.0.

[2] Odkaz ke stažení KUK 1.0 http://hdl.handle.net/11234/1-5821 a KUKY 1.0 http://hdl.handle.net/11234/1-5812.
[3] Odkaz na dokumentaci korpusu KUKY https://ufal.mff.cuni.cz/grants/ponk/kuky.
[4] https://ufal.mff.cuni.cz/udpipe/2, https://ufal.mff.cuni.cz/nametag/3
[5] https://arxiv.org/abs/1910.09513
[6] http://hdl.handle.net/11234/1-5225

Anotační schéma

Úřady, které se snaží zlepšovat svoji komunikaci s klienty, obvykle vyšlou jednoho či několik svých pracovníků na školení ve srozumitelném úředním psaní. Tito pracovníci pak sdílejí své zkušenosti s ostatními, nebo jsou dokonce pověřeni revidovat dokumenty svých kolegů. Někdy se v takovém případě uchovávají obě verze dokumentů pro další interní školení zaměstnanců.

Jistě existuje více způsobů, jak vyučovat úřední psaní a jak úspěšně revidovat dokumenty. V našem projektu jsme implementovali jednak metodu, kterou používají právní experti našeho aplikačního garanta, společnosti Frank Bold, jednak metodu Jany Šamánkové. J. Šamánková je advokátka s bohatou zkušeností s výukou i s managementem kvality dokumentů u různých zadavatelů (viz např. Šamánková a Kubíková, 2023). V předchozím výzkumu se autorský styl J. Šamánkové ukázal jako měřitelně srozumitelnější než běžné právní texty (Cinková, 2024). J. Šamánková svoji metodu dále rozvinula s Barborou Kubíkovou a spolu také anotovaly většinu dokumentů v korpusu KUKY 1.0.

Zatímco Frank Bold se zaměřuje na šablony právních úkonů pro obecnou veřejnost (např. jak založit spolek), J. Šamánková a B. Kubíková ve své praxi pracují spíše s argumentativními texty aplikujícími zákonnou normu na konkrétní situaci. Závěrem je zde stanovení práva či povinnosti, viny, či neviny, ale i pochybení, či nepochybení ze strany úřadu.

V obou praxích, Frank Bold a Šamánková & Kubíková, anotátor (1) nejprve důkladně pročte revidovaný text a označí nadbytečné a zcela nesrozumitelné části, které pak při revizi buď vynechá, nebo nahradí. Tuto anotaci označujeme Semafor relevance a používáme ji na všechny texty bez rozdílu; (2) v dalším kroku značí úryvky textu, sleduje, zda na sebe správně navazují různé typy úryvků podle jejich komunikačního účelu, a úryvky přeskupuje a přeformulovává. Tuto anotaci označujeme Anotace koherence. Texty z produkce Frank Bold mají zcela odlišný účel od textů, se kterými pracují Šamánková & Kubíková, a proto v Anotaci koherence používají odlišné sady značek. Korpus KUKY 1.0 je proto rozdělen na dva základní "žánry", každý se svou specifickou sadou značek uvedených v tabulce č. 1.

Metoda Šamánková & Kubíková vychází ze standardní právnické literatury (Gardner, 2007) a je založena na klasickém rétorickém postupu zvaném sylogismus. Sylogismus je velmi vhodný jak pro vysvětlování, tak pro přesvědčování. Skládá se ze tří výroků: obecný předpoklad, konkrétní předpoklad a závěr. Jeho rétorická úspěšnost závisí na tom, jak přirozeně se konkrétní předpoklad promítne na obecný předpoklad a jak přirozeně z obou vyplyne závěr. V ideálním případě by závěr čtenáři zdánlivě neměl přinést žádnou novou informaci, protože na totéž přišel sám po seznámení se s obecným a konkrétním předpokladem. V úředních textech je obecným předpokladem právní norma aplikovaná na daný případ, konkrétním předpokladem skutková podstata případu a závěrem je rozhodnutí úřadu. Úspěšné texty v obou předpokladech podávají pouze informace, které podporují čtenáře v samostatném vyvození následujícího závěru. V anotačním schématu sylogismu se používají značky uvedené v tabulce č. 1. U úspěšných dokumentů lze dokonce modelovat vztahy mezi označenými Pravidly, Příběhy a Závěry. To je

implementováno jako odkaz identifikátor daného úseku textu na identifikátor jiného úseku textu. Podrobnosti o anotačních schématech jsou uvedeny v dokumentaci korpusu KUKY na webové stránce projektu.[7]

Anotační nástroj Gloss

K anotaci Semafor relevance a Anotace koherence jsme použili nástroj Gloss, který vyvinul právní informatik Jaromír Šavelka pro své vlastní výzkumné projekty (např. Šavelka a kol., 2023) a který jsme jako vhodný anotační nástroj vybrali v prvním roce řešení projektu. J. Šavelka vytvořil uživatelské účty pro koordinátory anotace S. Cinkovou a M. Kuka a zaškolil je. Oba následně nastavili v Glossu anotační schémata a testovali s anotátory jednak anotační ergonomii uspořádání značek, jednak mezianotátorskou shodu, viz obrázky č. 2 a 3.

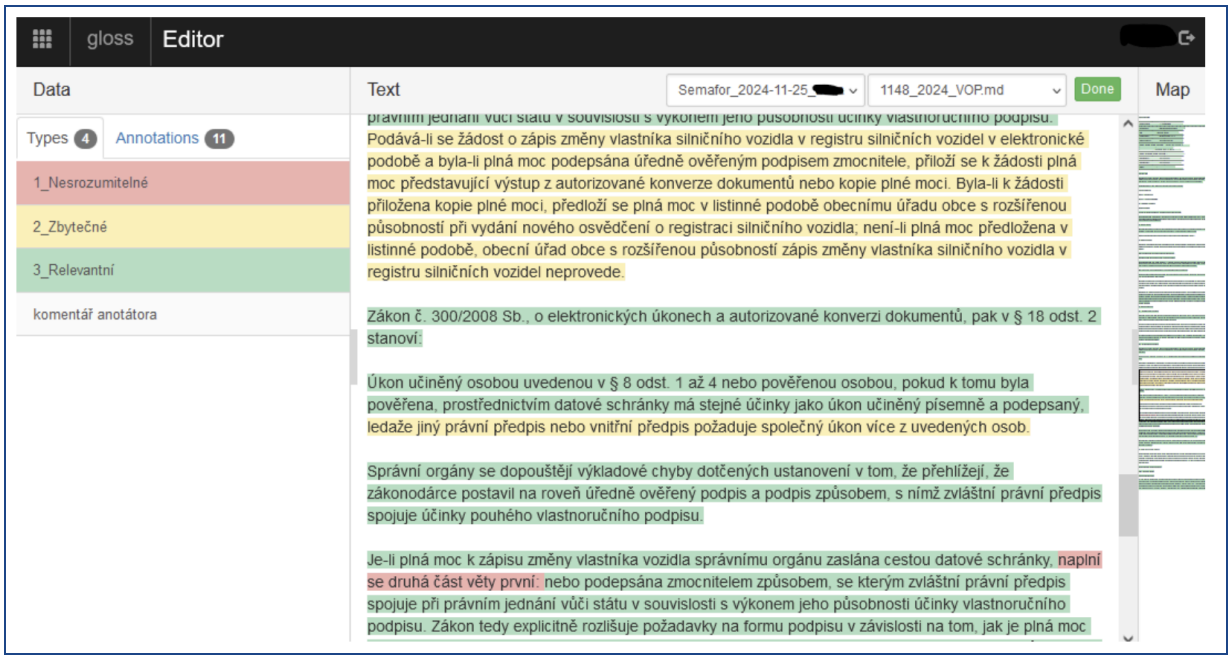
Anotační instrukce

Anotační instrukce pro anotátory pokrývaly instrukce, jak ovládat anotační nástroj Gloss, spíše než jak rozumět logice anotačních schémat, protože anotaci prováděli právní experti, kteří byli

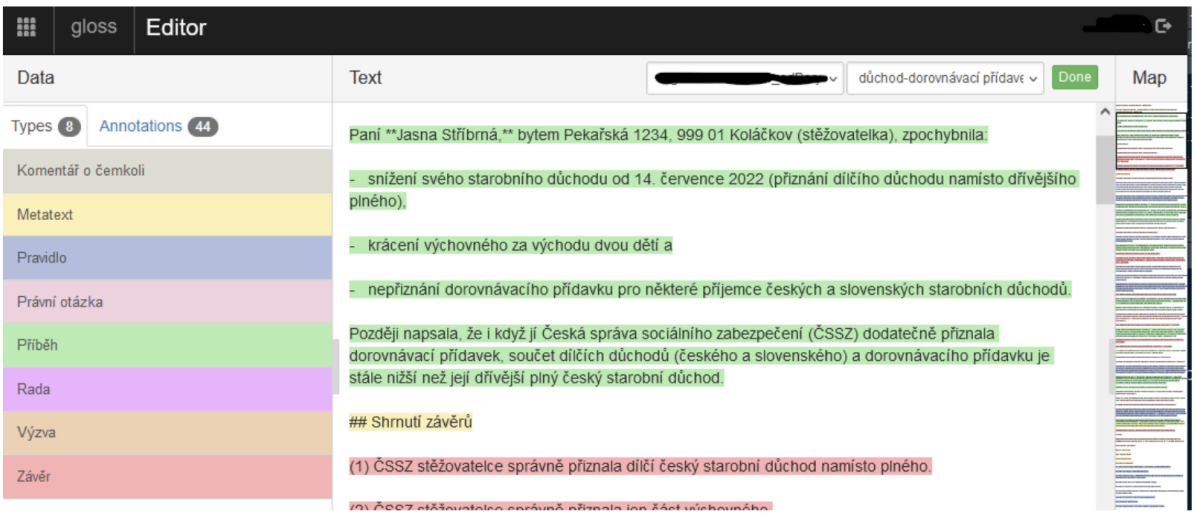
Frank Bold	Šamánková & Kubíková
situace	pravidlo (*obecný předpoklad*)
kontext	příběh (*konkrétní předpoklad*)
postup (*co můžete udělat*)	závěr
doporučení (*doporučení co dělat)*	právní otázka
proces (*co bude dělat úřad/protistrana*)	výzva
podmínky (*kdy něco můžete udělat*)	rada
odkazy (*na další podobné materiály Frank Bold*)	metatext (*procesní technikálie*)
prameny (*externí citace, především právní předpisy*)	
nezařaditelné	

Tabulka č 1. Anotační schema Anotace koherence

[7] https://ufal.mff.cuni.cz/grants/ponk/kuky



Obrázek 2: Anotace Semafor relevance v nástroji Gloss



Obrázek 3: Anotace koherence v nástroji Gloss

zvyklí s daným anotačním schématem pracovat analogově, často doslova s pastelkami a papírem.

Anotační proces

Vzájemné kalibrace anotátorů a měření mezianotátorské shody probíhala v několika iteracích od března 2024 a byla ukončena v posledním čtvrtletí roku 2024. Výsledkem je korpus KUKY 1.0.

1.3 Okruh Aplikace

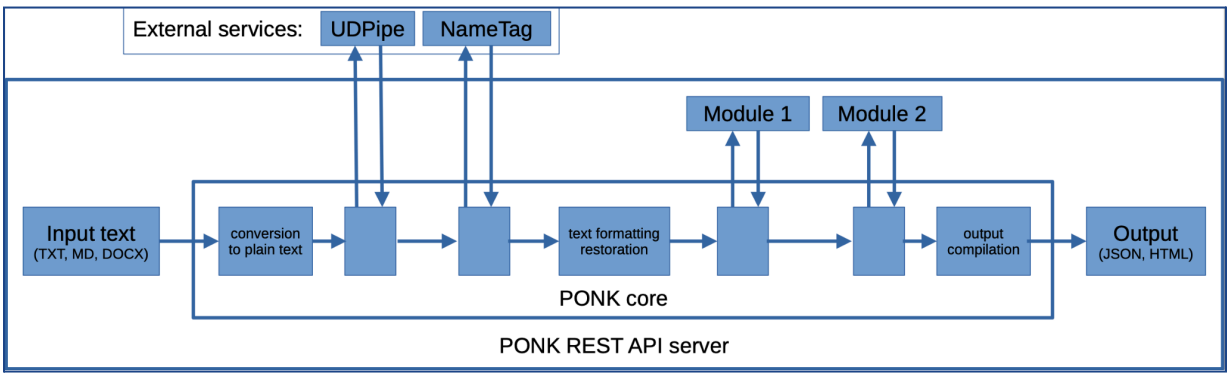
V rámci okruhu Anotace vznikla v roce 2024 první verze nástroje PONK (výsledek TQ01000526-V5) a pokračoval vývoj anonymizačního nástroje MasKIT, jehož první verze byla publikována v prvním roce řešení projektu (výsledek TQ01000526-V9).

PONK

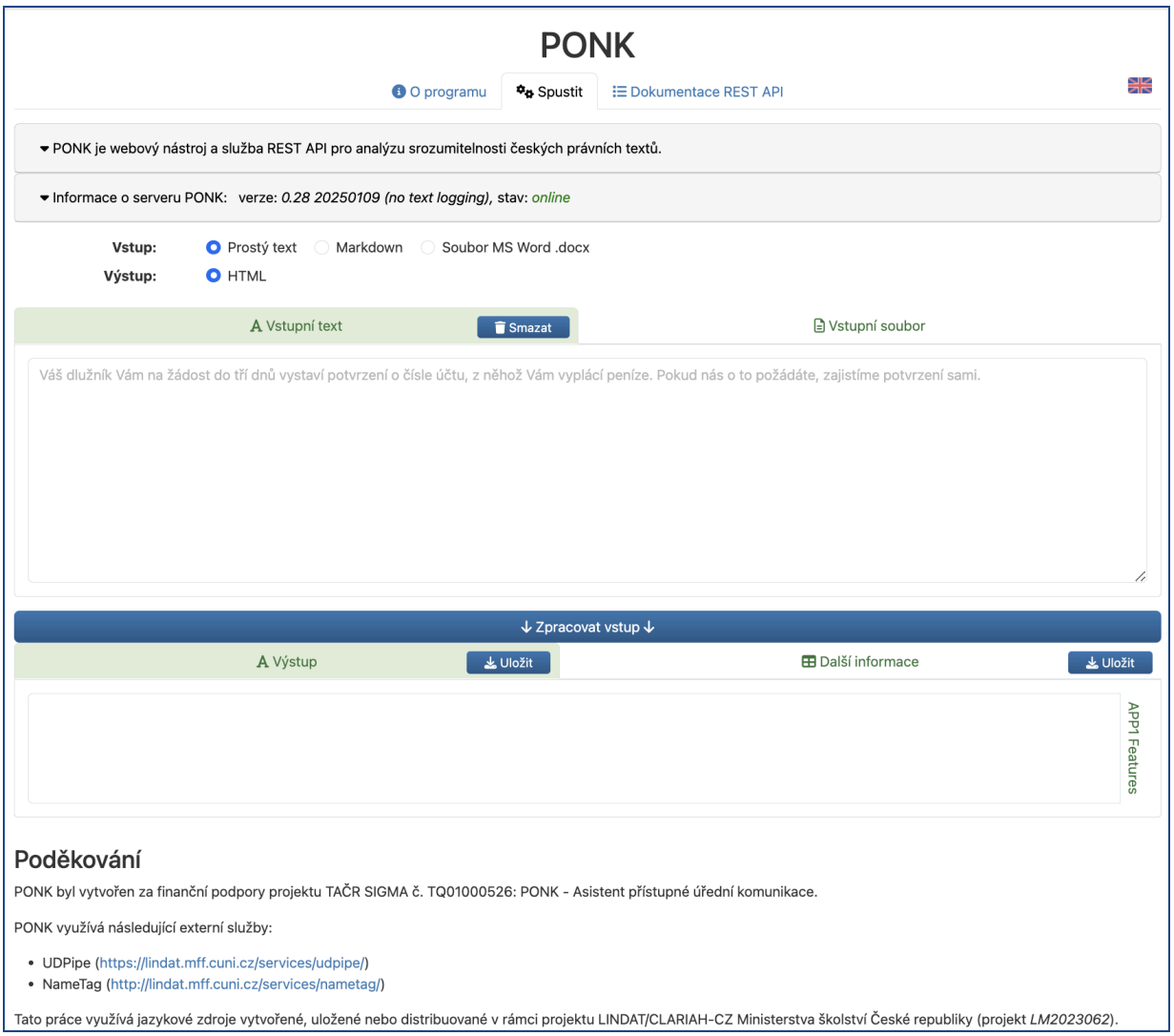
PONK 1.0 je navržen jako volně dostupná webová aplikace, jejímž cílem je identifikovat nesrozumitelné nebo obtížně srozumitelné části právního textu a doporučit uživateli (úředníkovi nebo právníkovi) přeformulování těchto částí. Detekce nesrozumitelných částí a určení stupně nesrozumitelnosti je založena na dvou metodách (modulech), a sice lexikální překvapení (modul 1, LEP) a lingvistická pravidla (module 2, LIP). Architektura PONK 1.0 je znázorněna na obrázku č. 4 a obrázek č. 5 ilustruje uživatelské rozhraní nástroje.

Modul lexikálního překvapení (LEP)

Lexikální překvapení se vztahuje k neočekávanosti nebo nepředvídatelnosti slova v daném kontextu, větě nebo textu. Jedná se o míru toho, jak překvapivé nebo neobvyklé je dané slovo na základě jazykových norem, pravděpodobností nebo očekávání čtenáře. Domníváme se, že úroveň lexikálního překvapení v textu má vliv na jeho srozumitelnost a že ji lze měřit pomocí pravděpodobnostního jazykového modelu natrénovaného na rozsáhlém množství textu. (Gehrmann a kol., 2019) vytvořili nástroj pro vizualizaci pravděpodobnosti textu zaměřeného na rozlišení textů generovaných umělou inteligencí od textů originálně sepsaných člověkem. My jsme jejich implementaci rozšířili a vytvořili modul nástroje PONK pro vizualizaci lexikálního překvapení jednotlivých slov jako indikátoru srozumitelnosti textu. Modul lexikálního překvapení nejdříve segmentuje text na dílčí slovní tokeny a následně každému tokenu přiřadí pravděpodobnost, kterou modul generuje ze slovníku vytvořeného dle použitého velkého jazykového modelu. Obrázek č. 6 demonstruje výstup modulu: úroveň překvapení je zobrazena pomocí barevného spektra, kde modrá představuje nízké překvapení, zelená střední úroveň překvapení a červená vysoké překvapení. Slova v tmavě modré barvě byla na základě



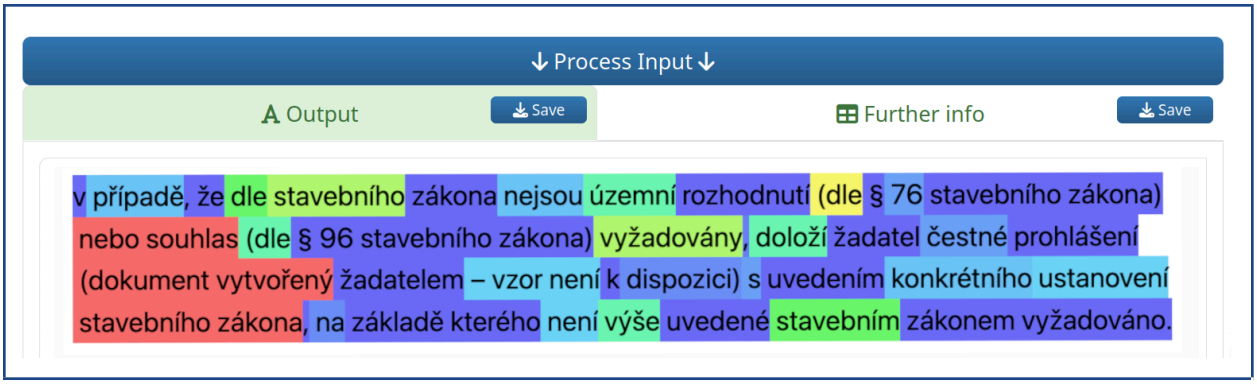
Obrázek 4: Architektura nástroje PONK 1.0



Obrázek 5: Uživatelské rozhraní aplikace PONK 1.0

předchozího kontextu zařazena mezi pět nejvýznamnějších kandidátů na jejich pozici v textu. Taková slova často patří k ustáleným výrazům nebo kolokacím. Rozšířené segmenty tmavě modrých segmentů mohou naznačovat redundanci, která snižuje srozumitelnost textu. Je třeba poznamenat, že míra lexikálního překvapení je vždy vyšší na začátku textu a stabilizuje se s přibývajícím textem.

Modul lexikálního překvapení vznikl v rámci bakalářské práce vedené Ivanou Kvapilíkovou, členkou řešitelského týmu. Práce je v době předkládání průběžné zprávy v oponentském řízení na MFF UK.



Obrázek 6: Vizualizace lexikálního překvapení v PONK 1.0

Modul lingvistických pravidel (LIP)

Jazykové vlastnosti, jako je délka věty, složitost slovní zásoby a syntax, hrají zásadní roli při určování čitelnosti, protože přímo ovlivňují porozumění a snadnost zpracování. Kratší věty obecně snižují kognitivní zátěž, což usnadňuje rychlejší zpracování a porozumění, zatímco delší, víceslovné věty mohou porozumění komplikovat, zejména u méně zdatných čtenářů. Podobně známá, vysoce frekventovaná slovní zásoba zvyšuje čitelnost tím, že činí texty přístupnějšími, zatímco složitá nebo málo frekventovaná slova mohou vést ke čtenářovu neporozumění nebo k rozladění (Hackemann, 2022), (Kleijn, 2018). Jednoduché syntaktické struktury zlepšují porozumění tím, že snižují kognitivní nároky, zatímco složité prvky, jako je dlouhá délka závislostí, mohou ztěžovat zpracování textů. Čtivost dále ovlivňují další jazykové faktory, včetně koherenčních značek, interpunkce a morfologických rysů. Například koherenční značky mohou v závislosti na jejich použití objasňovat vztahy mezi myšlenkami nebo komplikovat text. V konečném důsledku jsou účinky těchto jazykových rysů utvářeny kontextovými faktory, jako je čtenářova zdatnost, základní znalosti a zájem o téma, stejně jako účel a prostředí textu. Pochopení těchto interakcí je zásadní pro přesné hodnocení a zlepšování srozumitelnosti textů.

V modulu Lingvistických pravidel bylo implementováno 35 pravidel.[8] Ty jsou inspirované příručkami srozumitelné češtiny. (Šamánková a Kubíková, 2022) a (Šváb, 2021) jsou příručky určené úředníkům a odborníkům na právo, které se zaměřují především na srozumitelnost, jednoduchost a přístupnost. (Sgall a Panevová, 2014) je obecná příručka českého akademického psaní zaměřená na vědecké pracovníky a studenty bez lingvistického vzdělání. V (Chromý a Cehláková, 2023) je uvedeno více typů tzv. garden path sentences v češtině. K efektu garden path sentences dochází při čtení textu, kdy se ukáže, že struktura postavené věty je nesprávná (Trueswell a kol. 1993), tj. gramatická věta umožňuje alternativní rozbor, který se jeví jako správný dokud čtenář nenarazí na slovo, které k tomuto rozboru nelze připojit.

[8] https://github.com/VanaKraus/PONK-rules

Pravidla jsou implementována pomocí bloků Udapi[9] v programovacím jazyce Python. Kromě anotací pomocí nástrojů UDPipe a NameTag (viz výše) některá pravidla využívají nástroj MorphoDiTa k vyhledávání morfologických paradigmat lexémů.[10]

Pravidla pracují se syntaktickou reprezentací vět, se stromy generovanými nástrojem UDPipe. Pokud (pod)strom splňuje určitá kritéria, jsou slova ve (pod)stromě označena názvem příslušného pravidla a kategorií, která určuje, jakou roli má konkrétní slovo v rámci pravidla.

Každé pravidlo obsahuje stručný popis a doporučení, jak nesrozumitelný jev zlepšit. Pravidla jsou rozdělena do 5 kategorií podle povahy jevů, které zachycují:

- nejednoznačnost: konstrukce, které mohou vést k nejednoznačnosti (např. pravidla týkající se garden path sentences)
- shluky: jevy, které se ve větě objevují s takovou frekvencí, že mohou ohrozit srozumitelnost (např. mnoho záporů)
- fráze: specifické lexikální konstrukce (např. „slabikotvorná slova", zaměřené na vybraná lemmata).
- struktura: jevy související s gramatikou nebo strukturou věty (např. „predikát daleko ve větě").
- přijatelnost: jevy, které jsou ve standardní češtině považovány za negramatické (např. konstrukce jednak-jednak bez spojky)

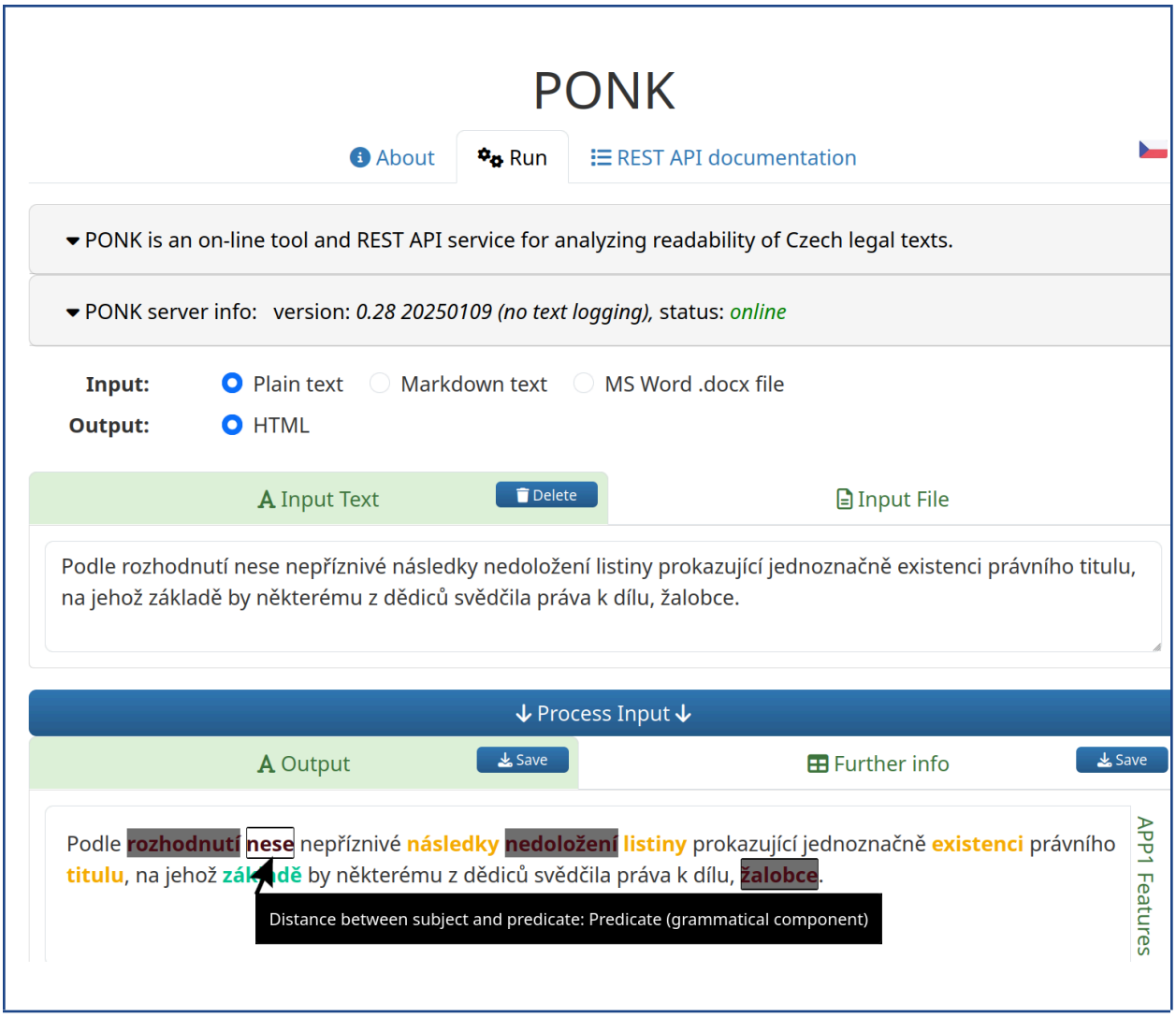
Kategorie jsou v grafickém rozhraní nástroje PONK vizualizovány barevně. Pokud je nějaké slovo "zasaženo" více pravidly, jeho pozadí je šedé, viz obrázek č. 7.

MasKIT

MasKIT je webový nástroj[11] a služba REST API pro anonymizaci českých textů se zaměřením na právní texty, jehož implementaci jsme zahájili a první verzi vydali v prvním roce řešení projektu (výsledek TQ01000526-V9). Referenčním dokumentem pro anonymizaci je Vyhláška č. 403/2022 Sb.[12]

MasKIT přijímá na vstupu český text (např. úřední rozhodnutí), v něm nalezne údaje (jména, adresy, rodná čísla, názvy firem, IČO, čísla pozemků atd.) a nahradí jejich výskyty náhodně zvolenými ekvivalenty, viz obrázek č. 8. Ke své činnosti využívá MasKIT UDPipe a NameTag. 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.[13] NameTag následně v textu vyhledá a klasifikuje jmenné entity (jména osob, adresy, názvy firem, geografická jména a mnoho dalších).

[9] https://ufal.mff.cuni.cz/morphodita
[10] https://ufal.mff.cuni.cz/morphodita
[11] https://quest.ms.mff.cuni.cz/maskit/
[12] Vyhláška o zveřejňování soudních rozhodnutí
[13] https://universaldependencies.org/



Obrázek 7: Vizualizace výstupů lingvistických pravidel v PONK 1.0

V roce 2024 byly odstraňovány chyby a laděny již existující vlastnosti MasKITu, např. při zpracování výrazu "hlavní město Praha" je nahrazeno nejen slovo "Praha", ale i skryto slovo "hlavní", kapitalizovaná slova jsou nahrazována slovy rovněž kapitalizovanými atd. Bylo přidáno zpracování dalších typů výrazů: data narození a úmrtí, čísla jednací, vládní agentury, politické instituce, vzdělávací a kulturní instituce, registrační značky vozidel. Naopak přestaly se anonymizovat názvy a místa soudů a jména soudců. Vylepšeno bylo programové (API) a uživatelské rozhraní, přidána byla i dvojjazyčnost rozhraní (angličtina a čeština). Kromě pseudonymizace (nahrazování podobnými slovy) byla přidána anonymizace (nahrazování třídami slov). Vývoj byl přesunut z repozitáře svn do repozitáře git.[14]

MasKIT byl členy řešitelského týmu testován na datasetu 23 soudních rozhodnutí nižších soudů (okresní, krajské) ve věcech civilněprávních i trestněprávních. Tento neanonymizovaný dataset

[14] https://github.com/ufal/maskit

byl poskytnut pro testování Ministerstvem spravedlnosti, Odborem elektronizace justice. Zahrnutá soudní rozhodnutí jsou plné texty prvoinstančních rozhodnutí. MasKIT byl testován skrz uživatelské rozhraní a manuálně evaluován z hlediska správnosti anonymizace rozhodnutí. Dále byl dále testován na právních dotazech z poradny Frank Bold Society, z.s. Takto bylo testováno přes 20 vybraných dotazů. Šlo o texty psané právními laiky a texty tak obsahovaly rovněž neformální či nepřesná označení.

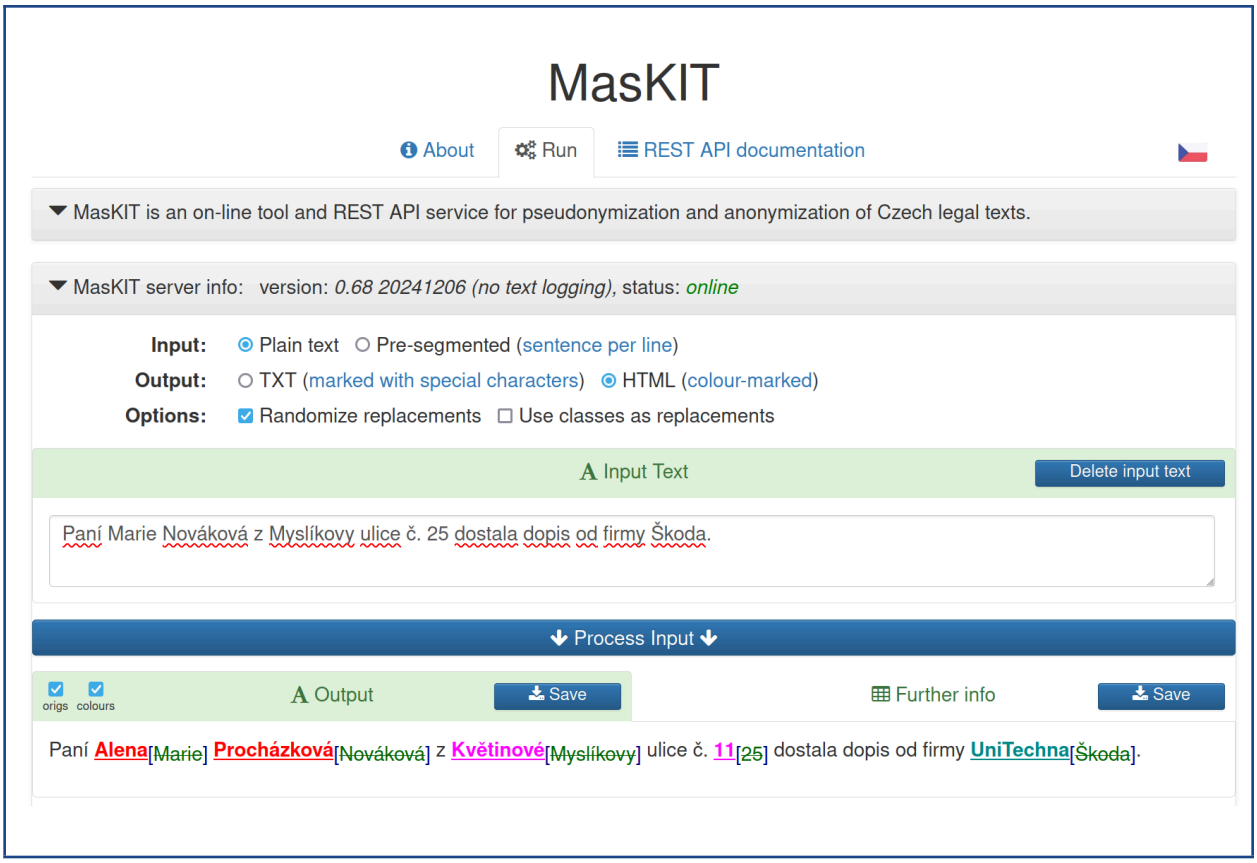
Statisticky významné vyhodnocení nástroje MasKIT vyžaduje více než 23 dokumentů. Obrátili jsme se na instituce (soudy a vybrané úřady), které mají ze zákona povinnost zveřejňovat příslušné dokumenty v anonymizované podobě, ohledně možnosti získat rovněž neanonymizovanou podobu dokumentů. Avšak bez úspěchu z důvodu právních rizik s tím spojených. Proto musíme evaluační data pro MasKIT vytvořit manuálně. V roce 2024 jsme v anotačním nástroji Brat připravili klasifikaci údajů pro jejich ruční anotaci v textech, viz obrázek č. 9. Anotace bude probíhat v roce 2025.

Oslovila nás společností Techniserv, která na objednávku Ministerstva spravedlnosti vytvořila a v roce 2024 spustila systém ISDRA, jehož součástí je i aplikace Anonymizér, umožňující anonymizaci soudních rozhodnutí pro celou soudní soustavu.[15] Dle jejich informací má Ministerstvo spravedlnosti zájem dále Anonymizér vylepšovat. Vývojáři společnosti Techniserv nyní testují, zda by MasKIT mohl být vhodným kandidátem na integraci do systému ISDRA pro další ulehčení práce soudních úředníků.

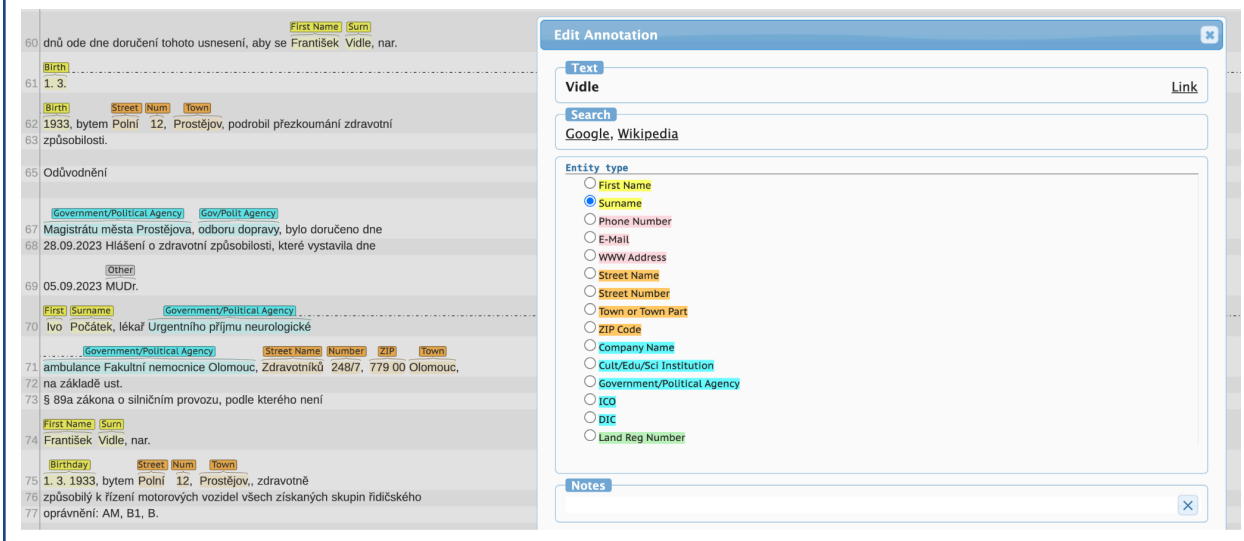
2. Prezentace projektu

- 10.-12.5.2024 — Ivan Kraus, Arnold Stanovský
  Prezentace modulu lingvistických pravidel
  Studentský workshop Žďárek, FF UK
- 28. 5. 2024 — Ivan Kraus, Arnold Stanovský
  Prezentace modulu lingvistických pravidel
  Deset minut pro studující při Dni češtiny, FF UK
- 13.9.2024 — Tereza Novotná
  Prezentace projektu PONK na konferenci České právo a informační technologie, Právnická fakulta MUNI

[15] https://smlouvy.gov.cz/smlouva/23742893



Obrázek 8: Anonymizační nástroj MasKIT



Obrázek 9: Ruční anonymizace textů v nástroji Brat

3. Ohlasy na projekt v mediích

Článek Tomáše Piky *Justiční úředníci si stěžují na nový anonymizér. ‚Chyby opravíme, vše je o zvyku,' kontruje ministerstvo* publikovaný na serveru iRozhlas se věnuje hodnocení anonymizéru od firmy Techniserv IT, vytvořeného v roce 2024 na objednávku Ministerstva spravedlnosti. V článku je zmíněn nástroj MasKIT, vyvinutý v rámci projektu PONK.

4. Způsob zapojení členů konsorcia

Ve druhém roce řešení projektu, leden-prosinec 2024, řešitelský tým splnil všechny tři naplánované výsledky

- KUKY 1.0 (SB - Specializovaná veřejná databáze, výsledek TQ01000526-V7)
- KUK 1.0 (SB - Specializovaná veřejná databáze, výsledek TQ01000526-V2)
- PONK 1.0 (Gfunk - Funkční vzorek, výsledek TQ01000526-V5)

Dílčí odpovědnosti členů řešitelského týmu byly následující:

Univerzita Karlova - Matematicko-fyzikální fakulta

Klíčové osoby

Barbora Vidová Hladká (řešitelka) byla zodpovědná za koordinaci řešení projektu, včetně přípravy průběžné zprávy za rok 2024; podílela se na publikaci korpusů KUKY 1.0 a KUK 1.0; participovala na metodologii testování modulů nástroje PONK; byla odpovědná za aktualizaci Plánu správy dat. Členkou řešitelského týmu je od 09/2023.

Silvie Cinková koordinovala anotaci korpusu KUKY 1.0. Vedla anotátorský tým, zajišťovala, aby anotace probíhaly systematicky a odpovídaly stanoveným instrukcím. Průběžně vyhodnocovala mezinotátorskou shodu a její analýzou reagovala zjemněním anotačních instrukcí. Vytvořila dokumentaci řešení anotační úlohy a korpus KUKY 1.0 připravila k publikaci v repozitáři LINDAT/CLARIAH-CZ. Členkou řešitelského týmu je od 09/2023.

Jiří Mírovský pokračoval ve vývoji nástroje MasKIT; konfiguroval anotační nástroj Brat k ruční anonymizaci textů pro statistickou evaluaci nástroje MasKIT; anotoval korpus KUK 1.0 pomocí nástrojů UDPipe a NameTag a publikoval ho v repozitáři LINDAT/CLARIAH-CZ; navrhl architekturu nástroje PONK a integroval do něho Modul lexikálního překvapení a Modul lingvistických pravidel. Členem řešitelského týmu je od 09/2023.

Ostatní osoby podílející se na řešení projektu

Ivana Kvapilíková, postdoc, koordinovala vývoj Modulu lexikálního překvapení v aplikaci PONK, který probíhal v rámci bakalářské práce na MFF UK, aktuálně přihlášené k obhajobě; participovala na integraci modulu do nástroje PONK; specifikovala instrukce pro ruční testování výstupů modulu LEP, které proběhne v 2025. Členkou řešitelského týmu je od 01/2024.

Tereza Novotná, postdoc, participovala na anotaci textů korpusu KUKY 1.0 a koordinovala práci na realizaci výsledku formou článku přihlášeného na konferenci ICAIL - AI and Law Conference 2025, Chicago; testovala funkční vzorek MasKIT a podílela se na formulaci anotačních instrukcí pro ruční anonymizaci textů; prezentovala projekt PONK na konferenci České právo a informační technologie pořádané Právnickou fakultou MUNI Brno. Členkou řešitelského týmu je od 09/2023.

Ivan Kraus, Arnold Stanovský, studenti FF UK, navrhli a implementovali Modul lingvistických pravidel (LIP) v aplikaci PONK. Modul prezentovali na akcích pořádaných FF UK.

Jana Šamánková, Barbora Kubíková, právní expertky, anotovaly argumentační strukturu v textech korpusu KUKY 1.0.

Lenka Fišerová, projektová administrátoka, zajišťovala projektovou správu projektu dle pravidel poskytovatele a řešitelské instituce. Členkou řešitelského týmu je od 09/2023.

Frank Bold Society

Klíčové osoby

Michal Kuk řešil zapojení aplikačního garanta do projektu. Průběžné konzultace a zapojení v rámci řešení projektu v roce 2024. Podílel se na testování rozhraní MasKIT. Členem realizačního týmu je od 09/2023. Připravoval a podílel se na anotaci části korpusu KUKY 1.0 v části normativních textů.

Ostatní osoby podílející se na řešení projektu

Právník právní poradny se v roce 2024 na projektu nepodílel.

Přemysl Pospíšil se na projektu podílel jako stážista právní poradny. Podílel se na anotaci normativních textů v korpusu KUKY 1.0.

5. Výsledky plánované na rok 2025

Výsledek TQ01000526-V3 (D - Stať ve sborníku) má termín dosažení 12/2025. Článek o dosud dosažených výsledcích jsme v lednu 2025 přihlásili na konferenci ICAIL - AI and Law Conference 2025, Chicago, USA. Dále připravujeme články o nástroji MasKIT a o anotaci argumentační struktury textu.

Výsledek TQ01000526-V4 (O - Ostatní výsledky) má termín dosažení 12/2025. Výsledek je souhrnná výzkumná zpráva o řešení projektu. Zaměříme se nejen na popis výsledků, ale i na dokumentaci kompletního řešení projektu, tj. popíšeme i rizika a nejistoty, se kterými jsme se potkali, a jak jsem přistoupili k jejich řešení. Zpráva bude určena všem organizacím, které mají srozumitelnost úředního psaní či právního ve svém programu. Zároveň bude určena autorům textů v jiných doménách, ve kterých je srozumitelnost rovněž velmi důležitá (např. lékařské znalecké posudky). Ve zprávě se totiž zaměříme i na dokumentaci reprodukovatelnosti.

Výsledek TQ01000526-V6 (Gfunk - Funkční vzorek) PONK 2.0 má termín dosažení 12/2025. Funkcionalita PONK 1.0 bude dále rozvíjena v rámci následujících dílčích aktivit tak, aby byl na konci projektu publikován PONK 2.0

- testování modulů LIP a LEP samostatně: statisticky významné hodnocení obou modulů vyžaduje ruční analýzu jejich výstupů. V prvním čtvrtletí 2025 bude provedeno ruční vyhodnocení modulů na vzorku textů z korpusu KUK 1.0.
- analýza korelace výstupů modulů LIP a LEP: jak zásahy pravidel korelují s mírou lexikálního překvapení. Formulujeme hypotézy, které budeme testovat na vzorku textů z KUK 1.0.
- implementace a testování nového modulu pro automatickou detekci argumentační struktury textů
- experimentování s modulem lingvistických pravidel v rámci prompt engineering nad velkými jazykovými moduly
- testování uživatelského rozhraní

Výsledek TQ01000526-V8 (W - Uspořádání workshopu) má termín realizace 6/2025. Na workshopu budeme prezentovat výsledky projektu. Demonstrujeme systematický postup od dat k nástroji, založený na nejmodernějších vědeckých přístupech, a zasadíme ho do národního a mezinárodního kontextu. Součástí budou praktické ukázky z průběhu testování jednotlivých modulů.

Literatura

- CINKOVÁ, Silvie, 2024. Linguistic Factors in the Readability of Czech Administrative and Legal Texts. In: To Understand Is to Be Free. Interdisciplinary Aspects of Comprehensibility and Understanding., pp. 303-325, Praesens Verlag, Vienna, Austria, ISBN 9783706912143.
- CURTOTTI, Michael, Eric McCREATH, 2013. A right to access implies a right to know: an open online platform for research on the readability of law. J. Open Access L., 1, 1.
- CURTOTTI, Michael, Eric McCREATH, Tom BRUCE, Sara FRUG, Wayne WEIBEL and Nicolas CEYNOWA, 2015. Machine learning for readability of legislative sentences. In Proceedings of the 15th International Conference on Artificial Intelligence and Law, 53–62.
- GARDNER, James A., Legal Argument, 2007. The Structure and Language of Effective Advocacy (Carolina Academic Press 2d ed. 2007).
- GEHRMANN, Sebastian, Hendrik STROBELT, and Alexander M. RUSH, 2019. GLTR: Statistical detection and visualization of generated text. CoRR, abs/1906.04043. http://arxiv.org/abs/1906.04043 arXiv: 1906.04043
- HACKEMANN, Timo, Lena HEINE and Dietmar HÖTTECKE, 2022. Challenging to read, easy to comprehend? effects of linguistic demands on secondary students' text comprehension in physics. International Journal of Science and Mathematics Education, 20, Suppl 1, 43–68.
- CHROMÝ, Jan and Markéta CEHÁKOVÁ, 2023. Diversity of garden-path structures - SPR. (Dec. 2023). doi: 10.17605/OSF.IO/KSTPE.
- KLEIJN, Suzanne, 2018. Clozing in on readability: How linguistic features affect and predict text comprehension and on-line processing. LOT.
- SEROTA, Michael, 2011. Intelligible justice. U. Miami L. Rev., 66, 649.
- SGALL, Petr and Jarmila PANEVOVÁ, 2014. Jak psát a jak nepsát česky. Karolinum,
- Prague. isbn: 978-80-246-2505-8.
- ŠAMÁNKOVÁ, Jana and Barbora KUBÍKOVÁ, 2023. Jak psát srozumitelné úřední texty. Příručka srozumitelného psaní pro úředníky. Veřejný ochránce práv, Brno, Czech Republic, ISBN 978-80-7631-088-9
- ŠAVELKA, Jaromír, 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 and Shiwanni JOHNSON, 2021. Lex Rosetta: transfer of predictive 9 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.

– ŠVÁB, Jakub, 2021. Jak psát, aby se to dalo číst. Příručka přístupného psaní. Leges, Prague. isbn: 978-80-7502-502-9.

– TRUESWELL, John C, Michael K TANENHAUS and Christopher KELLO. 1993. Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference From Garden-Paths. Journal of Experimental Psychology: Learning, Memory, and Cognition, 19, 3, 528–553.