

Projekt: Online služba pro vysoce kvalitní automatický překlad mezi slovanskými jazyky

JAZYKY:

Při řešení projektu jsme se zaměřovali především na překlad mezi češtinou, slovenštinou, polštinou a ruštinou. V poslední etapě projektu jsme se více věnovali i jihoslovanským jazykům, tedy slovinštině, chorvatštině, srbštině, bulharštině a makedonštině. Všechny překladače jsou plně funkční a nasazené na našich portálech ve více zemích, např.:

- <https://prekladac.lingea.cz/>
- <https://prekladac.lingea.sk/>
- <https://tlumacz.lingea.pl/>
- <https://prevajalnik.lingea.si/>
- <https://prevodilac.lingea.rs/>

Podarilo se nám připravit velmi kvalitní překladače nejen ze slovanských jazyků do češtiny a naopak, ale i mezi různými slovanskými jazyky navzájem.

První tabulka zobrazuje porovnání skóre BLEU mezi prvními verzemi našich překladačů a Google Překladačem na začátku grantu:

	BG	CZ	HR	MK	PL	RU	SI	SK	SR
BG	–	-3.4	0.7	-1.0	-4.2	-2.6	0.1	-2.9	1.5
CZ	2.7	–	0.3	-0.9	-2.0	-11.3	-1.3	-33.4	-0.4
HR	-1.2	0.7	–	-1.8	-1.1	-2.9	-1.0	0.8	1.0
MK	-1.7	0.2	0.2	–	-3.2	-4.4	-1.4	-1.4	0.9
PL	-0.4	-2.0	-0.2	-4.1	–	-4.2	-0.7	-3.8	-0.8
RU	-2.1	-8.4	1.3	-3.7	-3.7	–	1.2	-2.7	-0.6
SI	2.0	0.5	-1.0	-3.0	0.7	-1.9	–	0.2	0.1
SK	1.5	-28.0	-0.4	-4.0	-2.9	-1.4	-1.4	–	-0.1
SR	-1.2	0.6	0.9	-1.4	-2.4	-2.5	4.8	-6.4	–

Druhá tabulka ukazuje porovnání skóre BLEU mezi Lingea Translatorem a Google Překladačem na konci projektu. Z tabulky je patrné výrazné zlepšení, ve většině překladových směrů dosahujeme lepší skóre (kladné číslo) než Překladač Google (záporné číslo).

	BG	CZ	HR	MK	PL	RU	SI	SK	SR
BG	–	1.1	7.9	8.0	5.8	1.1	7.9	4.3	9.6
CZ	7.7	–	1.6	4.0	0.2	1.9	3.0	-3.9	5.4
HR	6.5	2.0	–	7.5	-0.2	-0.3	2.4	5.1	21.2
MK	6.0	3.8	8.7	–	-2.3	-3.5	5.8	4.5	9.5
PL	6.2	0.3	0.9	-2.2	–	-1.8	3.1	0.8	-0.1
RU	2.0	0.8	3.9	-1.6	0.2	–	5.3	4.1	-0.2
SI	8.1	4.1	2.7	4.4	3.3	2.8	–	6.3	13.5
SK	9.1	-1.2	5.2	4.3	-1.4	-0.8	6.5	–	10.8
SR	6.7	2.4	21.0	7.2	-1.8	-1.8	11.6	-2.6	–

BLEU (bilingual evaluation understudy) je algoritmus pro vyhodnocení kvality strojově přeloženého textu z jednoho přirozeného jazyka do druhého

FUNKCIONALITY:

1. Překlad neznačkováného textu, např. lokalizačních položek z databáze uživatele.
2. Překlad HTML stránek se zachováním formátování.
3. Překlad textů ve formátu XML se známou nebo uživatelsky definovanou sémantikou.
4. Překlad dokumentů (souborů) v dalších běžně používaných formátech

Funkcionality jsou implementovány pomocí API, jehož volání se parametrizuje tak, aby se dosáhlo jednotlivých cílů. Některé z funkcionalit jsou dostupné pomocí veřejně přístupných webových překladačů, ostatní jsou připraveny k integraci do informačních, datových či redakčních systémů našich potenciálních zákazníků.

Ukázky jednotlivých výstupů:

1. Překlad čistého textu podporujeme např. na portálu <https://slovnícek.cz>.
Volání, které přeloží např. lokalizační položku z databáze:
https://api.lingea.com/translate/?apiKey=apitran_i3saMdikk9zpLt1&locLang=cz&fromLang=cz&toLang=pl&text=Ru%C4%8Dn%C3%AD%20brzda
2. Překlad HTML textu např. z webové stránky podporuje portál <https://prekladac.lingea.cz>
3. Volání, které přeloží XML dokument lze demonstrovat na tomto jednoduchém příkladu:
<dokument><nadpis>Ukázkový XML dokument</nadpis><keywords>dokument, ukázka, XML</keywords><odstavec>Tento dokument je značkován pomocí XML. Značky oddělují jeho jednotlivé části a udávají jeho strukturu, takže je dobře strojově zpracovatelný.</odstavec><odstavec>Zároveň je tento ukázkový dokument dobře čitelný i pro člověka, protože značkování je jednoduché a intuitivní.</odstavec></dokument>
Příklad z češtiny do polštiny:
https://api.lingea.com/translate/?apiKey=apitran_i3saMdikk9zpLt1&locLang=cz&fromLang=cz&toLang=pl&fromFormat=structuralXML&toFormat=structuralXML&text=%3Cdokument%3E%3Cnadpis%3EUk%C3%A1zkov%C3%BD%20XML%20dokument%3C%3Cnadpis%3E%3Ckeywords%3Edokument,%20uk%C3%A1zka,%20XML%3C%3Ckeywords%3E%3Codstavec%3ETento%20dokument%20je%20zna%C4%8Dkov%C3%A1n%20pomoc%C3%AD%20XML.%20Zna%C4%8Dky%20odd%C4%9Bluj%C3%AD%20jeho%20jednotliv%C3%A9%20%C4%8D%C3%A1sti%20a%20ud%C3%A1vaj%C3%AD%20jeho%20strukturu,%20tak%C5%BEE%20je%20dob%C5%99e%20strojov%C4%9B%20zpracovateln%C3%BD.%3C%3Codstavec%3E%3Codstavec%3EZ%C3%A1roveň%20je%20tento%20dokument%20je%20zna%C3%BD%20dokument%20dob%C5%99e%20%C4%8Diteln%C3%BD%20i%20pro%20%C4%8Dlov%C4%9Bka,%20proto%C5%BEE%20zna%C4%8Dkov%C3%A1n%C3%AD%20je%20jednoduch%C3%A9%20a%20intuitivn%C3%AD.%3C%3Codstavec%3E%3C%3Cdokument%3E

Výsledek po zkopírování z JSON výsledku pak vypadá takto:

```
<dokument><nadpis>Przykładowy dokument XML</nadpis><keywords>Dokument, pokaz, XML</keywords><odstavec>Dokument ten jest oznaczany za pomocą XML. Kolczyki oddzielają jego poszczególne części i wskazują jego strukturę, dzięki czemu jest dobrze przetwarzany maszynowo.</odstavec><odstavec>Jednocześnie ten dokument pokazowy jest dobrze czytelny także dla człowieka, bo znakowanie jest proste i intuicyjne.</odstavec></dokument>
```

Programové rozhraní (API):

API je bodem, který umožňuje výsledky projektu reálně využívat v mnoha aplikacích. Běží na různých adresách na doméně api.lingea.com.

Konkrétně API překladače je dostupné na <https://api.lingea.com/translate/>. Pro jeho užití je potřeba mít založený takzvaný apiKey. Ten se dá založit v portálu pro správu API.

Off-line překladač

Lingea Translator Business Edition je překladač pro nasazení uvnitř firem, kde je možné ho používat při překladu a zpracování citlivých informací. Může používat tři typy uživatelských licencí s různými vlastnostmi a poskytuje administrační rozhraní, ve kterém ho lze velmi flexibilně konfigurovat. Administrátor může umožnit jak anonymní přístup, tak přihlašování jménem a heslem, a přidělovat licence:

1. pevné pro jednoho stálého uživatele,
2. plovoucí, kdy ji v každém okamžiku využívá nejvýše jeden uživatel a po jeho odhlášení je licence k dispozici ostatním uživatelům,
3. neomezené, kdy má zákazník k dispozici neomezený počet uživatelských licencí, omezen je jen celkový objem překladů za den

Na <https://ltbe-demo.lingea.cz/> je ukázka překladového rozhraní pro rusko-český a česko-ruský překladač.

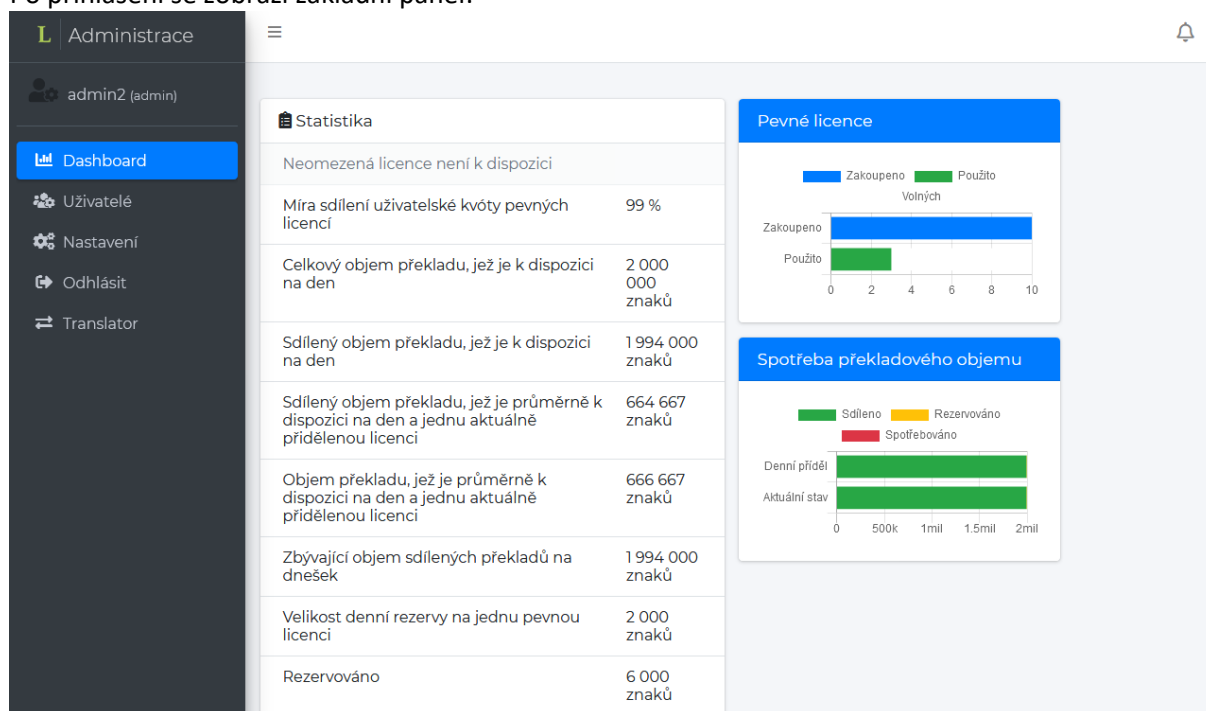
Pro vyzkoušení je možné se přihlásit pod těmito přihlašovacími údaji:

login: test1 heslo: kwRpad31

Možnosti administrace a monitorování jsou vidět na <https://ltbe-demo.lingea.cz/admin>, je možné se přihlásit takto:

login: admin2 heslo: kwRpad31

Po přihlášení se zobrazí základní panel:



Druhé okno umožňuje správu uživatelů:

Uživatelé

Login	Administrátor	Schválení	Licence	Pořadí	Rezervováno	Čas registrace	Poslední přihlášení	Poslední akce	IP adresa
admin1	Ano	Schválen	Pevná	1	2 000	12.05.2020 06:15	12.05.2020 10:04	12.05.2020 10:04	192.168.203.54
admin2		Schválen	Pevná	2	2 000	12.05.2020 06:16	12.05.2020 10:05	12.05.2020 10:28	192.168.203.41
test1	Ne	Schválen	Pevná	3	2 000	12.05.2020 06:22	12.05.2020 07:57	12.05.2020 07:59	192.168.203.241

(Položky: 1 - 3 z 3)

Plovoucí licence +

Neomezená licence +

Bez licence +

V dalším okně lze nastavovat licence:

Nastavení

- ☐ Použít zjednodušené přihlašování
- ☐ Automaticky schvalovat nové uživatele
- ☒ Přidělovat neomezené licence automaticky
- ☒ Přidělovat plovoucí licence automaticky
- ☒ Přidělovat pevné licence uživatelům automaticky
- ☐ Přidělovat pevné licence prohlížečům automaticky
- ☐ Uživatel může požádat o zapůjčení licencí

Míra sdílení uživatelské kvóty pevných licencí:

Hlavičky pro zjištění adresy:

Kontakt na administrátora:

Email na administrátora:

Pevné licence

Zakoupeno: 10, Použito: 3, Volných: 7

Spotřeba překladového objemu

Sdíleno: 2mil, Rezervováno: 1.5mil, Spotřebováno: 0.5mil

Podpůrné jazykové nástroje:

Nekvalitní vstupní text je často hlavním důvodem chybného překladu. Některé opravy lze provést automaticky a nástroje pro ně jsme do API integrovali v rámci grantu. Jedná se především o doplnění diakritiky a opravy překlepů.

Na portálu <https://korektor.lingea.cz/> je možné vyzkoušet tyto nástroje pro jazyky řešené v rámci projektu, tedy češtinu, slovenštinu, chorvatštinu, polštinu, ruštinu, slovinštinu a srbštinu.

E-shop pro prodej on-line jazykových služeb a monitorování jejich využití:

Všechny překladové a korekturní nástroje a služby si mohou uživatelé z řad firemní klientely pořídit na e-shopu společnosti Lingea. V současnosti probíhá integrace všech typů produktů (knih, on-line i

off-line) do nového společného e-shopu včetně přechodu na novou platební bránu. Část zaměřená na monitorování a nastavování API překladačů a dalších jazykových nástrojů vypadá následovně:

apiKey	Translate	Add diacritics	Spell check	Detect language	Manage configuration	Monitor usage
all_12345687890asdfghjklqwertyuiopzxcvbnm	☒	☒	☒	☒	☒	☒
work_12345687890asdfghjklqwertyuiopzxcvbnm	☒	☒	☒	☒	☐	☐
admin_12345687890asdfghjklqwertyuiopzxcvbnm	☐	☐	☐	☐	☒	☒

KORPUSY:

Pro trénování překladačů jsou zapotřebí paralelní korpusy, tedy takové, které obsahují stejné texty v obou jazycích daného jazykového páru vedle sebe. Paralelní data jsou získávána z veřejně dostupných zdrojů. Procházejí několika stupni filtrování, které se provádí pro každý větný pár zvlášť. Nejprve se vyřadí páry, které jsou duplicitní a také takové, kde jsou věty v obou jazycích identické. Dále se odstraňují různé netisknutelné znaky, tvrdé mezery a různé varianty interpunkce se normalizují na běžně používané alternativy. Poté přichází na řadu detekce jazyka, při které model založený na neuronových sítích určuje pravděpodobnost příslušnosti věty do jazyka na základě posloupností znaků ve větě.

Pokud obě věty z páru náleží do předpokládaných jazyků, dostanou se do další fáze filtrování, ve které se používají překladové slovníky Lingea. Tímto způsobem filtrujeme především data, kde předpokládáme hodně chyb a nesprávně zarovnaných vět. Jsou vyřazeny takové větné páry, kde jedna věta obsahuje příliš mnoho výrazů, které podle slovníku nejsou možným překladem žádného výrazu z druhé věty, nebo takové, které obsahují slova, která se ve slovníku vůbec nenachází.

Dále se překlady filtrují na základě délky pomocí heuristik a jsou vyřazeny takové věty, ve kterých je na jedné straně více než čtyřnásobný počet slov, než na druhé.

V další fázi jsou k odstranění nekvalitních párů použity už existující překladové modely, natrénované na menších, ale kvalitnějších datech. Nejprve se hodnotí pravděpodobnost toho, že by model pro překlad ze zdrojového do cílového jazyka přeložil zdrojovou větu druhou větou z páru. Stejným způsobem se spočítá pravděpodobnost překladu v opačném směru. Pokud jsou pravděpodobnosti dostatečně podobné a dostatečně vysoké, větný pár je zachován pro další zpracování. Poslední stupeň filtrování má za cíl odstranit takové páry vět, které jsou sice správně přeložené, ale nejde o příliš přínosná data pro trénování a naopak mohou uškodit. Většinou jde o specifikace produktů z e-shopů, menu webových stránek, seznamy apod.

Kvalitních paralelních dat je většinou velmi omezené množství, které nedostačuje k trénování kvalitních překladových modelů. Proto je potřeba vyrobit si umělá trénovací data. K tomu používáme

texty v cílovém jazyce, které pomocí vlastního nástroje stahujeme z webových stránek. Jednojazyčných textů je na internetu mnohem více, než paralelních dat. Texty se pročistí a přeloží překladačem do zdrojového jazyka. Například pro zlepšení překladu z češtiny do polštiny stáhneme texty z polských webových stránek (nejčastěji zpravodajské servery), přeložíme je už existujícím překladačem do češtiny a přidáme k česko-polským paralelním trénovacím datům. Pro zlepšení efektivity takového trénování se do zdrojové strany přidává „šum“, např. se vypouští nebo nahrazují některá slova, nebo se mění jejich pořadí ve větě. Lze použít i opačný přístup, tedy stáhnout data ve zdrojovém jazyce a přeložit je do cílového, ale ten nemívá tak dobré výsledky. Tímto způsobem jsme vylepšili překladače ve směrech, které byly pro nás nejvíce důležité - všechny kombinace mezi češtinou, polštinou, ruštinou a slovenštinou.

U vícejazyčných modelů je výhodné, pokud model sdílí znalosti naučené u jednotlivých směrů, tedy že např. trénováním překladu z ruštiny do češtiny selepší i překlad z bulharštiny do češtiny. Abychom toto sdílení usnadnili, experimentovali jsme s přepisem textů v jazycích psaných azbukou do latinky. To modelu umožní slova s podobným základem stejně segmentovat a tedy sdílet reprezentaci podobných slov mezi jazyky psanými různou abecedou. Úskalím transliterace je, že s výjimkou srbštiny není jednoznačná a při převodu z azbuky do latinky se může ztratit informace. Snažili jsme se tento problém omezit vývojem vlastní jednotné transliterace pro všechny slovanské jazyky (po experimentech s různými transliteracemi na základě národních nebo ISO norem). Jako nejlepší řešení se ukázalo transliterovat pouze zdrojové věty trénovacích dat a cílové věty ponechat v původní abecedě - odpadá pak problém zpětné transliterace do azbuky, která opět nemusí být jednoznačná.

Velikosti paralelních korpusů ukazuje následující tabulka, uvedené údaje jsou v tisících řádků.

	BG	CZ	HR	MK	PL	RU	SI	SK
CZ	3 850							
HR	1 114	963						
MK	548	357	453					
PL	4 079	5 834	1 086	417				
RU	1 270	1 595	699	491	3 247			
SI	3 103	4 350	695	215	4 377	558		
SK	3 458	5 198	741	214	2 131	557	4 203	
SR	761	624	770	547	786	929	300	3 640

V tabulkách jsme použili tyto zkratky: BG-bulharština, CZ-čeština, HR-chorvatština, MK-makedonština, PL-polština, RU-ruština, SI-slovinština, SK-slovenština, SR-srbština.