



Ústav formální a aplikované lingvistiky

Thesis Areas

2026

Overview

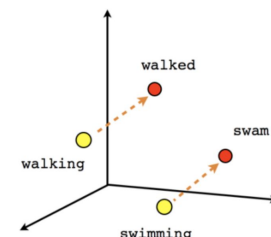
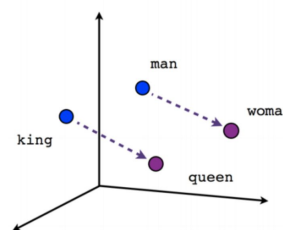


- Choose your mix of:
 - Natural language processing
 - Deep learning
 - Computational linguistics
 - Machine translation
 - +Music!
- Diverse concentration of [experts](#), competition winners, ...
Solid % of theses are good enough research to be published.
- Many international projects & industrial cooperation
- Internship support (     universities)
- Cluster: >140 GPUs (>5TB RAM),
>2000 CPUs (>32TB RAM)
- Try our web services at lindat.cz



Pestrost jazyka, modelování prostoru vět

- Jednu větu lze přeformulovat tisíci způsoby a zachovat význam.
 - Včera přišel Novák pozdě. František Novák dorazil jsem včera se zpožděním. F. Novák má další pozdní příchod. Franta zas přišel pozdě.
- Ale i nepatrnou úpravou lze význam **zásadně změnit**.
 - Včera **ne**přišel Novák pozdě. **J.** Novák má další pozdní příchod.
- Neuronové sítě umějí větu převést do mnohazměrného spojitého prostoru....
- ...dají se v tomto prostoru najít vektory odpovídající “normálním lingvistickým úpravám”?
 - Nováka kritizovali. -(drsně)-> Nováka seřvali.
 - Nováka kritizovali. -(otočit význam)-> Nováka chválili. Nováka nekritizovali.
 - Nováka kritizovali. -(neurčitě)-> Někomu tam něco vytýkali.
 - **Sebráno 7k vět:** COSTRA 1.1, změřeny základní větné embeddingy
- Například: **lepší srozumitelnost právnických a byrokratických textů!**



Modelling Language Evolution

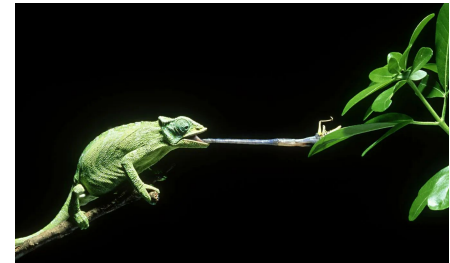
- Word migration across languages

- Words are frequently borrowed from one language to another
- Evolution of vocabulary cannot be captured by a human, because
 - zillions of morphemes in thousands of languages
 - all kinds of sound changes along the way, different alphabets...
- Not just etymology for individual languages...
- We want to **model evolution of language as a global phenomenon!**



- Competition in Natural Languages

- Languages are just different solutions to identical communication needs.
- Like in living nature, possible solutions of particular “tasks” compete
- Examples:
 - Fixed word order vs. richer morphology?
 - Original vs. loan-word to express a new concept?



- In both cases:

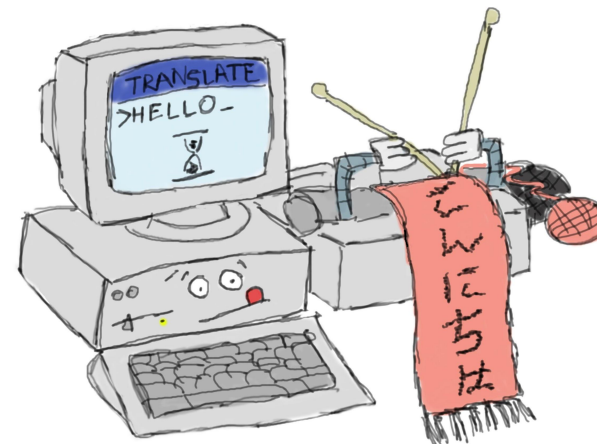
- **Timely topics:** massively multilingual data available only in the last few years
- Rich experience with multi-lingual data at UFAL MFF UK
- Inspiration could be taken from **mathematical models in biology** (predator-prey models, dynamic equilibria in microbio ecosystems, horizontal gene transfer...)
- The last missing piece is **YOU** – a mathematically gifted machine learning practitioner

Strojový překlad (Neural Machine Translation)

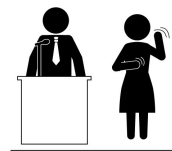
- Náš anglicko-český překladač [překonal v přesnosti profesionální překladatele](#) ([ověřte](#) si sami), ale **jak pokrýt mnoho jazyků?**

Kolik ztratíme překladem přes Aj?

- Natrénovat přímé překlady a porovnat s překladem přes pivotní jazyk a multilinguálním LLM
- OpenEuroLLM
 - Otevřená evropská 37-jazyčná LLMka! Široké možnosti...
 - Hlavní projekt evropské suverenity v oblasti umělé inteligence



- Dvojazyčný Google Doc
 - textový editor pro dvousloupcové dokumenty (smlouvy ap.) s NMT
- Dialekty & jazykové kulturní dědictví! Překlad do ostravštiny/hantecu...



Součástky pro strojové tlumočení

Strojový překlad mluvené řeči lze zjednodušeně vidět jako posloupnost:

- ASR (rozpoznávání mluvené řeči) + MT (strojový překlad)



V mnoha ohledech [jde o těžší úkol](#) (viz [elitr.eu](#)):

- Lidé nemluví ve větách => **Vylepšit segmentaci**, aby brala v úvahu prozodii a hledala členění na “myšlenky”.
- V přednáškách jsou speciální termíny, vlastní jména => Vytvořit systém pro **okamžitou doménovou adaptaci pro ASR**.
 - specializované rozpoznávání řeči, které bude **pouze zachytávat klíčové termíny**, které **daný řečník těsně před přednáškou cvičně přečte**.
- Lidé mluví rychleji, než čtou => Trénovat **sumarizaci mluvené řeči**, podobně jako to dělají tlumočníci.
- Někdy je čas ručně živé rozpoznávání korigovat => Vytvořit **živý editor přepisu a překladu mluvené řeči**.
 - Důraz na klávesové zkratky, automatické učení se oprav, “souhru” s korektorem.

Součástky pro strojové tlumočení

- Offline editor ASR

- Spojit zvuk a text do jedné UI komponenty
- Využívat confidence ASR systému, např. barvit slova
- Důraz na jednoduchost a rychlost oprav; automatické návrhy



Generování uměleckého textu

- Automatické generování **poezie, scénářů, povídek**
 - Automatic generation of poetry, scripts, stories
 - GPT apod.
 - Measuring and/or modelling creativity
- ...a vyhodnocování!
- Vizualizace slovních reprezentací
 - visualisation of word embeddings
- Info: <http://ufal.cz/rudolf-rosa/projekty>



Co se děje na sociálních sítích

- Projekt **Infoveillance** detekuje informační anomálie na sociálních sítích
 - kombinujeme jazykové technologie a grafové algoritmy
 - <https://ufal.mff.cuni.cz/grants/infoveillance>

“Tady nežijí neplatí daně nemá volba být plátna podle mého nepřípustné nežijí v ČR”

- Jak si poradí model trénovaný na novinových článcích s tímto příspěvkem?
- Moc dobře ne.
- Proto musíme
 - adaptovat modely na žánr sociálních sítí
 - trénovat nové modely nebo ladit existující
 - naučit modely rozlišovat emoci, zkratku, manipulaci, ...

Question Answering from Health Records

Understanding a **hospital discharge report** or other patient health records is typically very difficult for the patient....

Question Answering systems could read such reports and generate (lay-language) answers to user's questions in natural language (part of an EU project)

Důvod přijetí : iCMP pons l.sin.

Průběh hospitalizace :

Pacientka přijata cestou KCC z interního odd. Český Krumlov, kde zjištěny náhle vzniklé parestezie dx. končetin bez jiných neurologických příznaků, vstupně dekomp. hypertenze. Při příjezdu na KCC už pouze odeznívající parestezie v oblasti pravého ústního koutku, při vertikalizaci ale neschopna chůze. Na CT ictovém protokolu bez známek akutní ischemie či krvácení, doplněna MR mozku, kde nález akutní ischemie premediálně v pontu vlevo, bez demarkace na FLAIR sekvenci. Pro potencioálně invalidizující deficit i při NIHSS 0b. indikována IVT. Vstupně lab. bez hrubé patologie. Pacientka monitorována, EKG se SR bez záchytu Fis. Kontrolní CT mozku bez demarkace čerstvých ložisek ischemie. Během LTV pacientka samostatně chodící s dopomocí 1 fr. berle. Stabilní, KP komp., schopná dimise. Při dimisi orientovaná, afebrilní, bez neurologického deficitu.

Závěr: 28.12. akutní iCMP v pontu l.sin., CTA neg., MRI DWI/FLAIR +/-, NIHSS 0, ale invalidizující deficit - neschopna samostatného stoje a chůze - st.p. podání IVT, pre-mRS 0, TOAST 5 - při dimisi NIHSS 0, mRS 0

Dg: Základní dg : I633 minor stroke LICA
Ostatní dg : E785 - Hyperlipidemie NS
U5300 NIHSS 0

Doporučená medikace :

Anopyrin 100mg 0-1-0
Trombex 75mg 0-1-0 na 1 měsíc
Controloc 40mg 1-0-0 na 1 měsíc
Torvacard 80mg 0-0-1

chronicky:

Nebilet 5mg 1/2-0-0
Cipralax 10 1,5-0-0

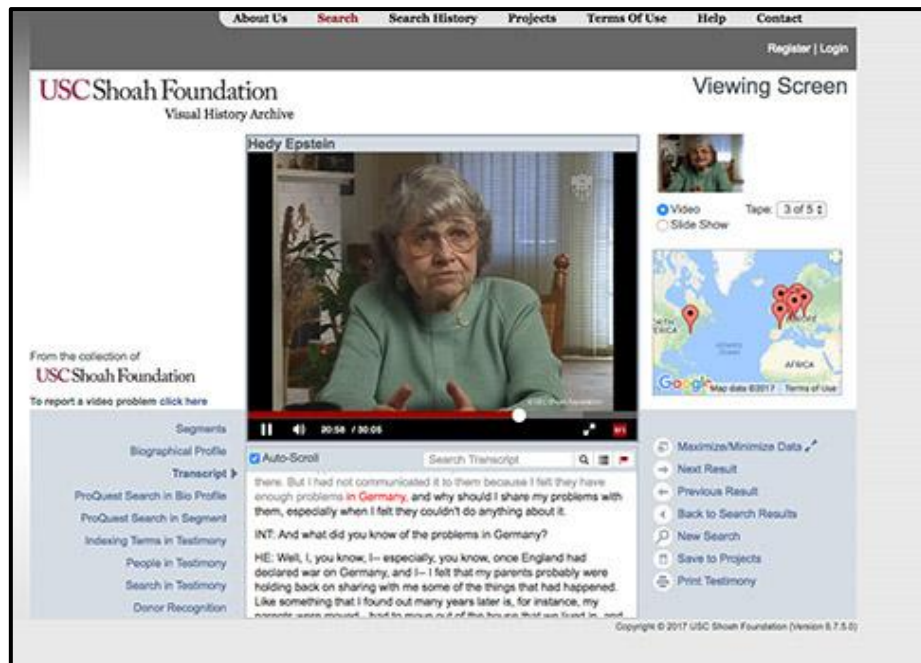
Q: What was my blood pressure?

A: 175/95 mmHg

Working with Historical Texts

Textual historical materials exist in very large amounts (digitized diaries, testimonies) and languages. Accessing such materials (browsing/searching) can be improved by Semantic Tagging where the data is linked to elements from an ontology and other information.

(the work will be done as a part of an EU project, **a huge dataset for training neural models is available**)



credit <https://sfi.usc.edu/>

Topic: *pre-school education*

Place: *Terezin, 50°30'40", 14°9'2"*

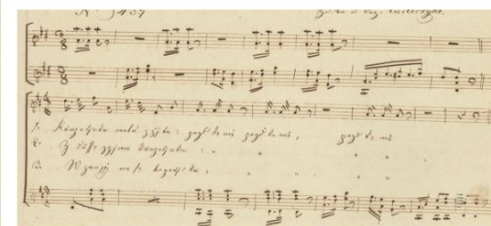
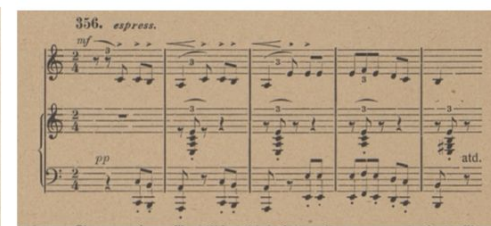
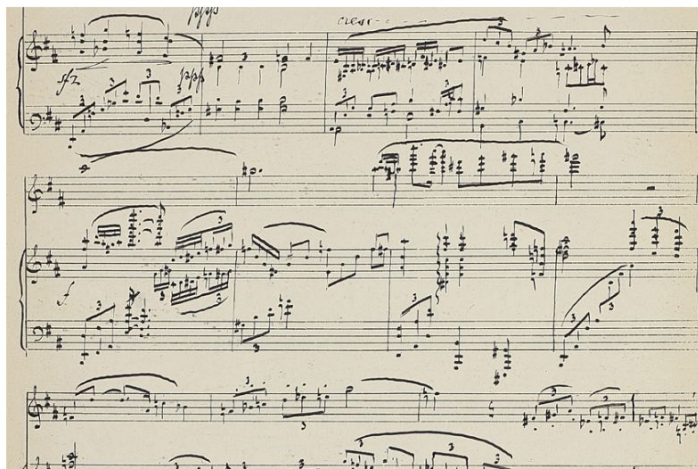
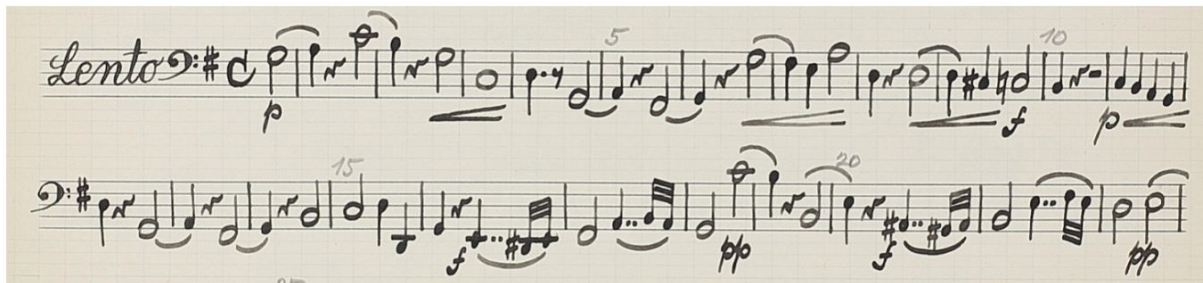
People: *Jan Novák, Petr Pospíšil*

Date/Period: *Dec 1942*

Organization: ...

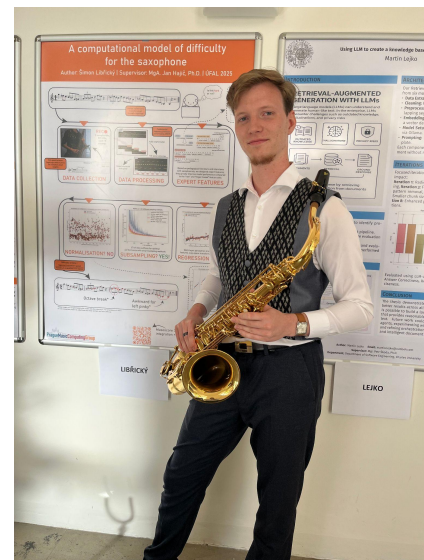
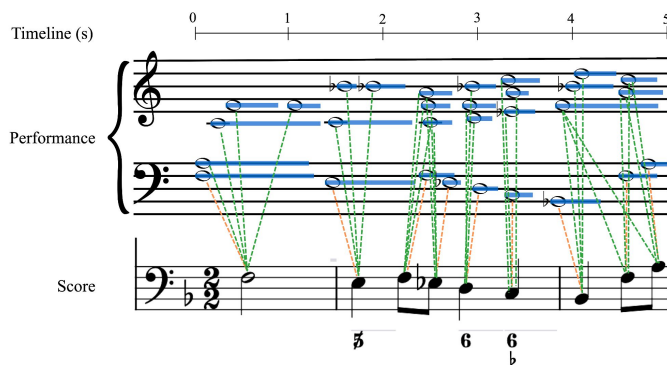
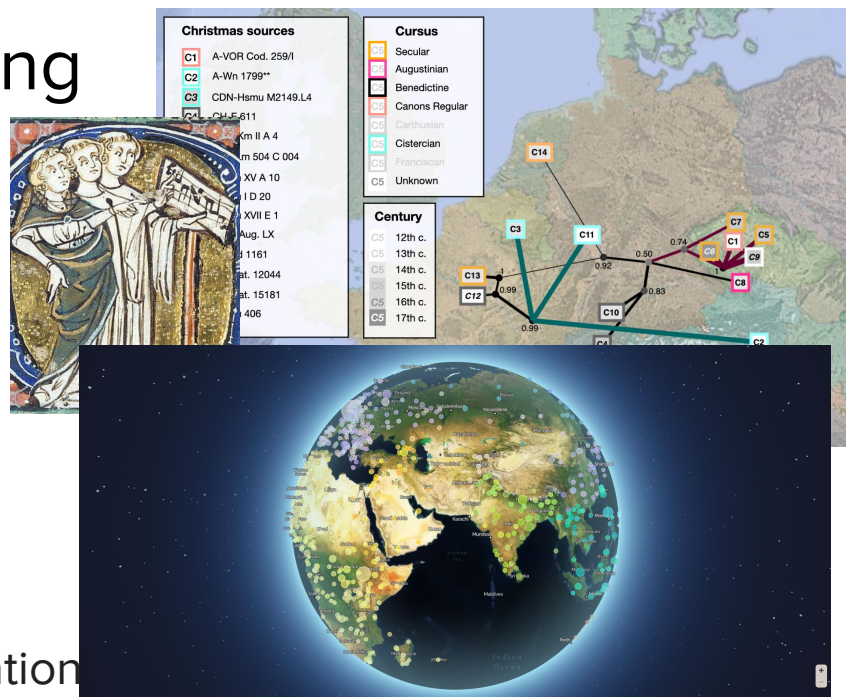
Optical Music Recognition

How to computationally read musical notation? (Hard problem. Vision LLMs are... bad at this.)



Computational Music Processing

- Cracking secrets of medieval music that people still sing using **bioinformatics** (10^6 records from medieval manuscripts)
- Comparing musics from around the world (e.g., Kurdish vs. Czech wedding songs...)
- Graph neural networks to model improvisation
- Creating tools that make it easier to make music (together) (e.g., difficulty model for saxophone arrangements...)



Feel free to write to us at thesis@ufal.mff.cuni.cz