Ústav formální a aplikované lingvistiky

Bachelor's Thesis Topics Proposals

at ÚFAL

February 2023

ÚFAL overview

- choose your mix of:
  - Natural language processing
  - Deep learning
  - Computational linguistics
  - Dialogue systems
  - Machine translation
- amazing concentration of experts,
  shared task winners (CoNLL, WMT,...)
- many international projects and industrial cooperation
- internship support (.. universities)
- cluster: 100 GPUs (>1TB RAM),
  >2000 CPUs (>32TB RAM)
- try our web services at lindat.cz

Jak porozumět 138 jazykům?

- Lidský jazyk: spousta výjimek a nejednoznačností
- Jeho fungování lze „natrénovat“ z velkých dat

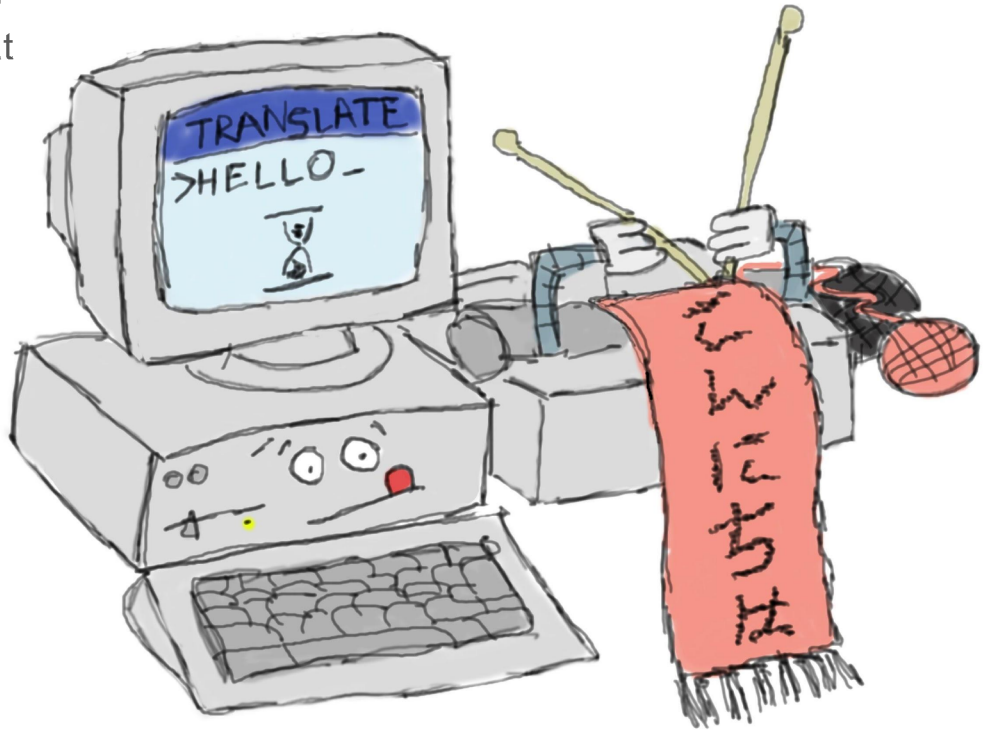
- Máme jazyková [data pro 138 jazyků](), od staré řečtiny až po vietnamštinu
- Práce s mnohojazyčnými daty = celá řada témat na bakalářky i diplomky!

- Věděli jste např., že může být jeden minulý čas pro včerejšek a jiný pro vzdálenější minulost?

Prospective supervisor: Daniel Zeman <zeman@ufal.mff.cuni.cz>

Strojový překlad (Neural Machine Translation, NMT)

- 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
- Překlad webových stránek
  - zachovat markup, přeložit obsah
- NMT pro reformátování
  - automaticky upravit např. bibliografické záznamy podle vzoru
- Dvojjazyčný Google Doc
  - textový editor pro dvousloupcové dokumenty (smlouvy ap.) s NMT



- Využití NMT pro překlad do ostravštiny/hantecu

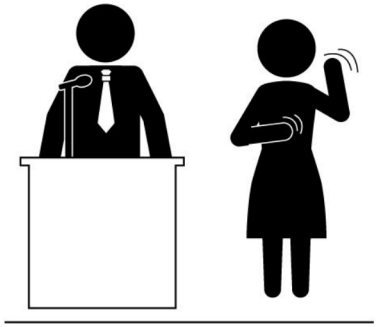
Prospective supervisor: Martin Popel, Ondřej Bojar, Jindřich Helcl

Součástky pro strojové tlumočení

Strojový překlad mluvené řeči lze zjednodušeně vidět jako posloupnost:

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

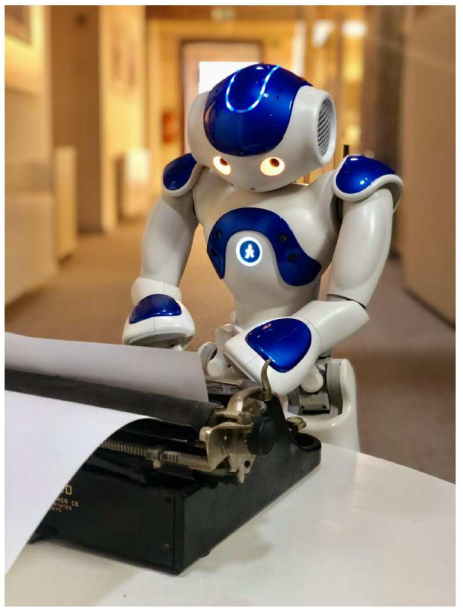
V mnoha ohledech jde o těžší úkol (viz elitr.eu):



- 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 pr**
  - 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.
- Offline editor ASR
  - Spojiť zvuk a text do jednej UI komponenty
  - Využívať confidence ASR systému, napr. farbiť slová
  - Dôraz na jednoduchosť a rýchlosť opráv; automatické návrhy

Prospective supervisor: Ondřej Bojar <bojar@ufal.mff.cuni.cz>, Peter Polák <polak@ufal.mff.cuni.cz>

Generování textu / Generating text



- automatické generování poezie, scénářů, povídek
  - automatic generation of poetry, scripts, stories
  - GPT apod.

Interpretace neuronových sítí / Interpreting neural networks

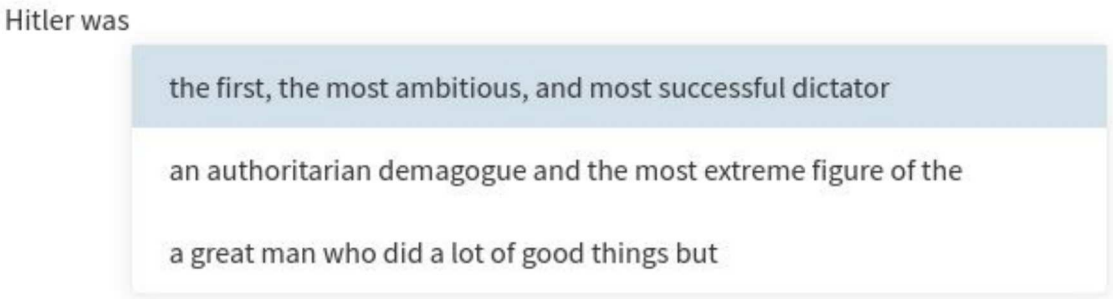
- vizualizace slovních embedinků
  - visualisation of word embeddings

- Jsem na hranici kapacity, mohu vzít už max 1 studenta, možná!
  - I am at the limits of my capacity, I can take max 1 more student and only maybe!
- Info: http://ufal.cz/rudolf-rosa/projekty

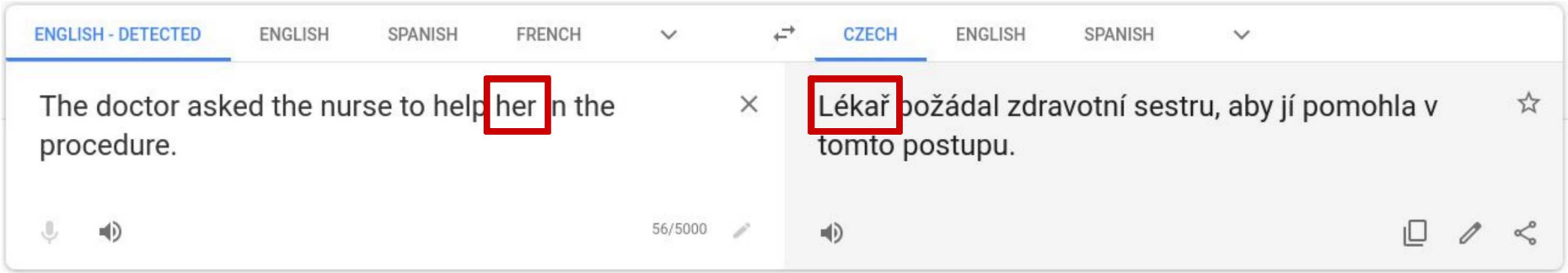
Prospective supervisor: Rudolf Rosa <rosa@ufal.mff.cuni.cz>

Stereotypy v neuronových sítích

GPT-2
Model pro generování textu



- Neuronové sítě potřebují pro svoje trénování obrovské množství textu, které se dá najít jenom na internetu
- Internet je plný divných textů od divných lidí
- Modely mohou z dat pochytit nežádoucí stereotypy, které mohou negativně ovlivnit fungování aplikací
- Anglické modely jsou rasistické a sexistické… co ty trénované na češtině?



Prospective supervisor: Jindřich Libovický <libovicky@ufal.mff.cuni.cz>

Jazyk rozdělené společnosti

- Metody pro tzv. neřízený strojový překlad se dají použít pro překlad mezi skupinami obyvatel
- Můžeme vytvořit slovníček pojmů, které se objevují v souvislosti s polarizujícími tématy
- Nebo dokonce překládat do jazyka antivaxerů…

Čemu by odpovídala slova jako *rouškař* nebo *odmítač* v češtině roku 2019?

Jak řekne stejnou informaci ekologický aktivista a jak pro-ruský troll?

Americký slovníček: dvojice slov používaných ve stejném významu <demokraty, republikány>

Category	Misaligned pairs
Political entities	⟨democrats, republicans⟩, ⟨blue, red⟩, ⟨dem, republican⟩, ⟨gop, democrats⟩, ⟨schumer, mcconnell⟩
News entities	⟨fox, cnn⟩, ⟨tapper, carlson⟩, ⟨tapper, hannity⟩, ⟨tucker, cuomo⟩, ⟨lemon, hannity⟩
Derogatory	⟨boarder, border⟩, [49], ⟨republicunts, democraps⟩, ⟨maddow, madcow⟩, ⟨democrats, demoncrats⟩, ⟨cuomo, shithead⟩, ⟨obama, obummer⟩, ⟨schiff, schitt⟩, ⟨spanky, trump⟩
(Near) synonyms	⟨lmao, lol⟩, ⟨stupidest, dumbest⟩, ⟨wh, whitehouse⟩, ⟨sociopath, psychopath⟩, ⟨favor, favour⟩, ⟨hahaha, hahahah⟩, ⟨hillary, hrc⟩, ⟨congresswoman, pocahontas⟩
Spelling errors	⟨melanie, melania⟩, ⟨kellyann, kellyanne⟩, ⟨hillary, hilary⟩, ⟨avenatti, avenati⟩
Ideological	⟨protests, riots⟩, ⟨progressives, socialists⟩, ⟨socialists, communists⟩, ⟨bigotry, paranoia⟩, ⟨liberals, conservatives⟩, ⟨communism, nazism⟩, ⟨commies, fascists⟩, ⟨liberalism, conservatism⟩, ⟨racism, supremacy⟩

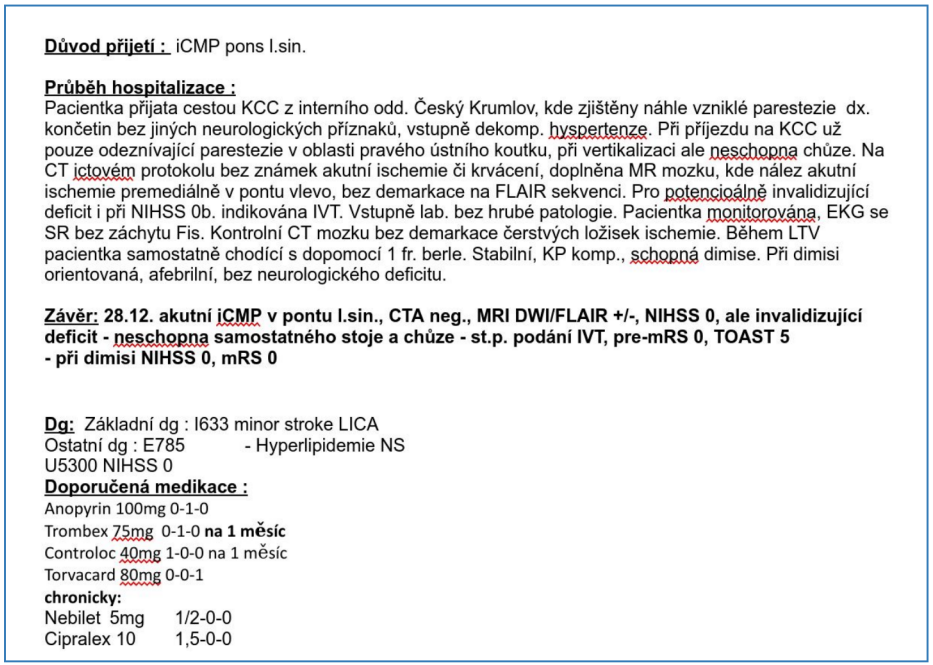
Zdroj: https://arxiv.org/pdf/2010.02339.pdf

Prospective supervisor: Jindřich Libovický <libovicky@ufal.mff.cuni.cz>

Question Answering from Health Records

Understanding a hospital discharge reports or other patient health records is typically very difficult for a layperson. A possible solution is a neural Question Answering systems that reads such a report and generate (lay-language) answers to user's questions in natural language.

(the work will be done as a 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. hyspertenze. 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
Cipralex 10 1,5-0-0

Q: What was my blood pressure?

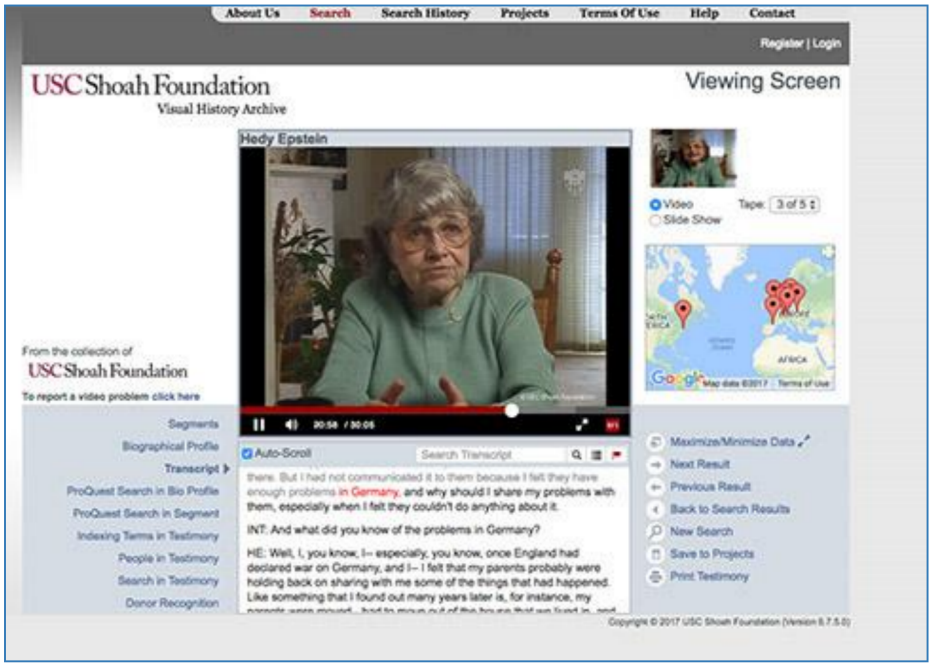
A: 175/95 mmHg

Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>

Semantic Tagging of 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***

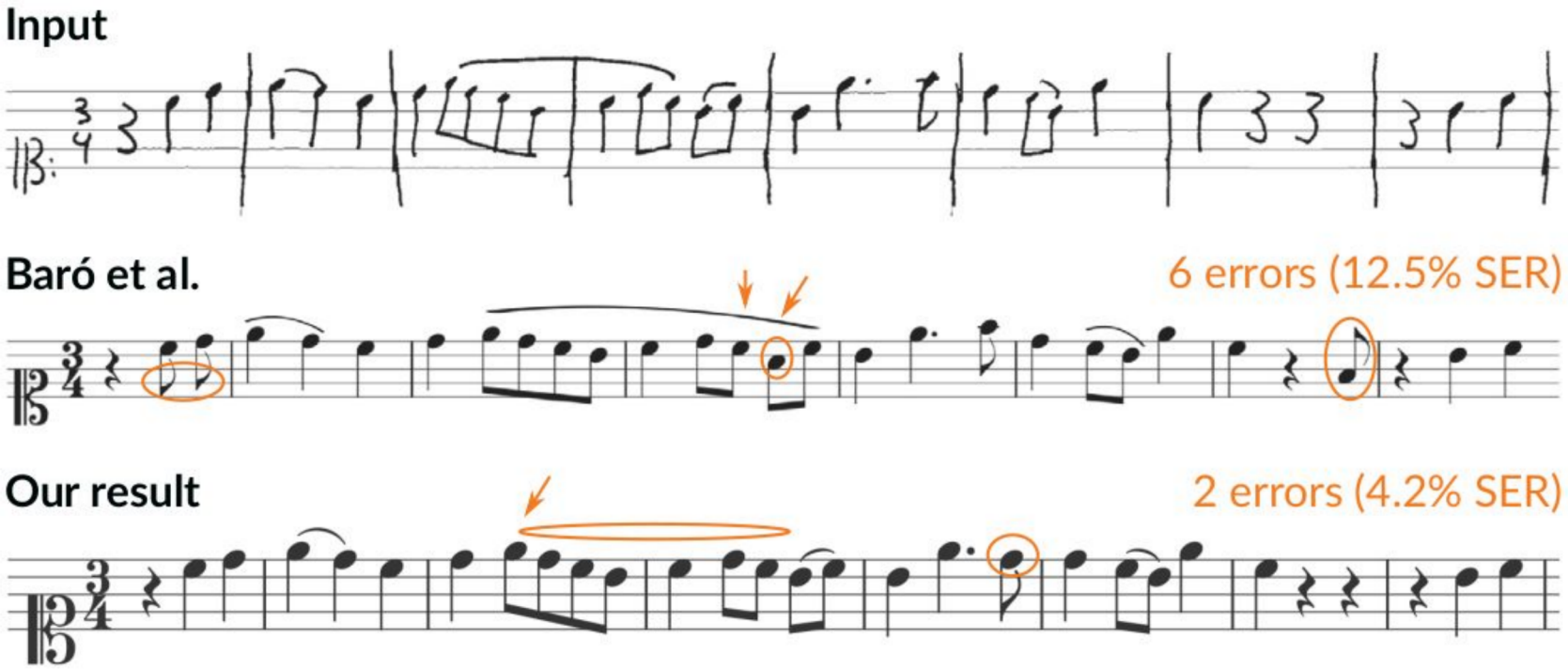
Date/Period: ***Dec 1942***

Organization: …

Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>

Optical Music Recognition

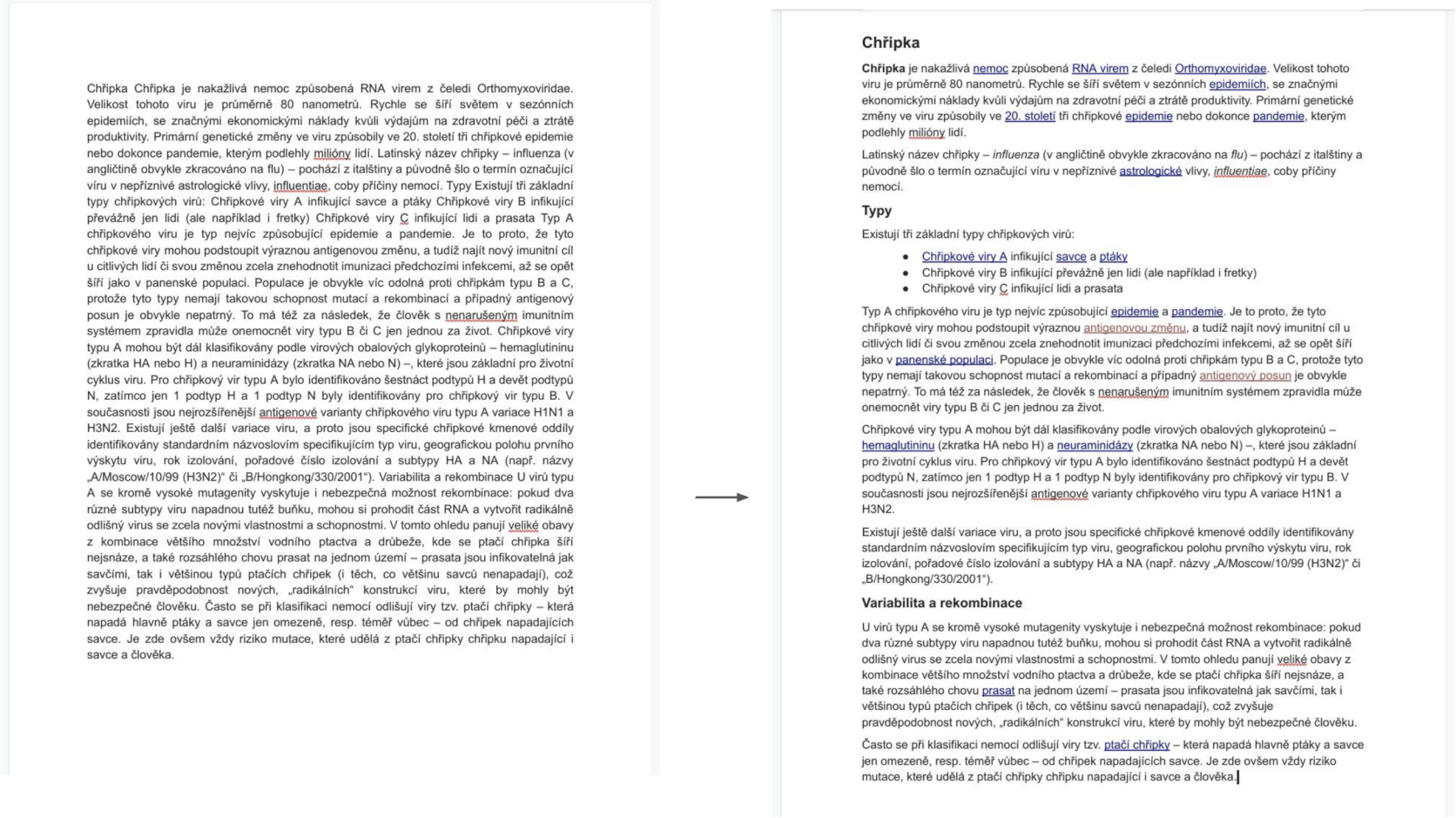
How to computationally read musical notation in documents (printed/handwritten)



Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>

Text Structuring

Automatic splitting of written text into **meaningful units**, such as **paragraphs**, **itemized lists, sections,** etc. eventually generating titles and headlines.



Chřipka Chřipka je nakažlivá nemoc způsobená RNA virem z čeledi Orthomyxoviridae. Velikost tohoto viru je průměrně 80 nanometrů. Rychle se šíří světem v sezónních epidemiích, se značnými ekonomickými náklady kvůli výdajům na zdravotní péči a ztrátě produktivity. Primární genetické změny ve viru způsobily ve 20. století tři chřipkové epidemie nebo dokonce pandemie, kterým podlehly miliony lidí. Latinský název chřipky – influenza (v angličtině obvykle zkracováno na flu) – pochází z italštiny a původně šlo o termín označující víru v nepříznivé astrologické vlivy, influentiae, coby příčiny nemocí. Typy Existují tři základní typy chřipkových virů: Chřipkové viry A infikující savce a ptáky Chřipkové viry B infikující převážně jen lidi (ale například i fretky) Chřipkové viry C infikující lidi a prasata Typ A chřipkového viru je typ nejvíc způsobující epidemie a pandemie. Je to proto, že tyto chřipkové viry mohou podstoupit výraznou antigenovou změnu, a tudíž najít nový imunitní cíl u citlivých lidí či svou změnou zcela znehodnotit imunizaci předchozími infekcemi, až se opět šíří jako v panenské populaci. Populace je obvykle víc odolná proti chřipkám typu B a C, protože tyto typy nemají takovou schopnost mutací a rekombinací a případný antigenový posun je obvykle nepatrný. To má též za následek, že člověk s nenarušeným imunitním systémem zpravidla může onemocnět viry typu B či C jen jednou za život. Chřipkové viry typu A mohou být dál klasifikovány podle virových obalových glykoproteinů – hemaglutininu (zkratka HA nebo H) a neuraminidázy (zkratka NA nebo N) –, které jsou základní pro životní cyklus viru. Pro chřipkový vir typu A bylo identifikováno šestnáct podtypů H a devět podtypů N, zatímco jen 1 podtyp H a 1 podtyp N byly identifikovány pro chřipkový vir typu B. V současnosti jsou nejrozšířenější antigenové varianty chřipkového viru typu A variace H1N1 a H3N2. Existují ještě další variace viru, a proto jsou specifické chřipkové kmenové oddíly identifikovány standardním názvoslovím specifikujícím typ viru, geografickou polohu prvního výskytu viru, rok izolování, pořadové číslo izolování a subtypy HA a NA (např. názvy „A/Moscow/10/99 (H3N2)" či „B/Hongkong/330/2001"). Variabilita a rekombinace U virů typu A se kromě vysoké mutagenity vyskytuje i nebezpečná možnost rekombinace: pokud dva různé subtypy viru napadnou tutéž buňku, mohou si prohodit část RNA a vytvořit radikálně odlišný virus se zcela novými vlastnostmi a schopnostmi. V tomto ohledu panují veliké obavy z kombinace většího množství vodního ptactva a drůbeže, kde se ptačí chřipka šíří nejsnáze, a také rozsáhlého chovu prasat na jednom území – prasata jsou infikovatelná jak savčími, tak i většinou typů ptačích chřipek (i těch, co většinu savců nenapadají), což zvyšuje pravděpodobnost nových, „radikálních" konstrukcí viru, které by mohly být nebezpečné člověku. Často se při klasifikaci nemocí odlišují viry tzv. ptačí chřipky – která napadá hlavně ptáky a savce jen omezeně, resp. téměř vůbec – od chřipek napadajících savce. Je zde ovšem vždy riziko mutace, které udělá z ptačí chřipky chřipku napadající i savce a člověka.

Chřipka

Chřipka je nakažlivá nemoc způsobená RNA virem z čeledi Orthomyxoviridae. Velikost tohoto viru je průměrně 80 nanometrů. Rychle se šíří světem v sezónních epidemiích, se značnými ekonomickými náklady kvůli výdajům na zdravotní péči a ztrátě produktivity. Primární genetické změny ve viru způsobily ve 20. století tři chřipkové epidemie nebo dokonce pandemie, kterým podlehly miliony lidí.

Latinský název chřipky – *influenza* (v angličtině obvykle zkracováno na *flu*) – pochází z italštiny a původně šlo o termín označující víru v nepříznivé astrologické vlivy, *influentiae*, coby příčiny nemocí.

Typy

Existují tři základní typy chřipkových virů:

- Chřipkové viry A infikující savce a ptáky
- Chřipkové viry B infikující převážně jen lidi (ale například i fretky)
- Chřipkové viry C infikující lidi a prasata

Typ A chřipkového viru je typ nejvíc způsobující epidemie a pandemie. Je to proto, že tyto chřipkové viry mohou podstoupit výraznou antigenovou změnu, a tudíž najít nový imunitní cíl u citlivých lidí či svou změnou zcela znehodnotit imunizaci předchozími infekcemi, až se opět šíří jako v panenské populaci. Populace je obvykle víc odolná proti chřipkám typu B a C, protože tyto typy nemají takovou schopnost mutací a rekombinací a případný antigenový posun je obvykle nepatrný. To má též za následek, že člověk s nenarušeným imunitním systémem zpravidla může onemocnět viry typu B či C jen jednou za život.

Chřipkové viry typu A mohou být dál klasifikovány podle virových obalových glykoproteinů – hemaglutininu (zkratka HA nebo H) a neuraminidázy (zkratka NA nebo N) –, které jsou základní pro životní cyklus viru. Pro chřipkový vir typu A bylo identifikováno šestnáct podtypů H a devět podtypů N, zatímco jen 1 podtyp H a 1 podtyp N byly identifikovány pro chřipkový vir typu B. V současnosti jsou nejrozšířenější antigenové varianty chřipkového viru typu A variace H1N1 a H3N2.

Existují ještě další variace viru, a proto jsou specifické chřipkové kmenové oddíly identifikovány standardním názvoslovím specifikujícím typ viru, geografickou polohu prvního výskytu viru, rok izolování, pořadové číslo izolování a subtypy HA a NA (např. názvy „A/Moscow/10/99 (H3N2)" či „B/Hongkong/330/2001").

Variabilita a rekombinace

U virů typu A se kromě vysoké mutagenity vyskytuje i nebezpečná možnost rekombinace: pokud dva různé subtypy viru napadnou tutéž buňku, mohou si prohodit část RNA a vytvořit radikálně odlišný virus se zcela novými vlastnostmi a schopnostmi. V tomto ohledu panují veliké obavy z kombinace většího množství vodního ptactva a drůbeže, kde se ptačí chřipka šíří nejsnáze, a také rozsáhlého chovu prasat na jednom území – prasata jsou infikovatelná jak savčími, tak i většinou typů ptačích chřipek (i těch, co většinu savců nenapadají), což zvyšuje pravděpodobnost nových, „radikálních" konstrukcí viru, které by mohly být nebezpečné člověku.

Často se při klasifikaci nemocí odlišují viry tzv. ptačí chřipky – která napadá hlavně ptáky a savce jen omezeně, resp. téměř vůbec – od chřipek napadajících savce. Je zde ovšem vždy riziko mutace, které udělá z ptačí chřipky chřipku napadající i savce a člověka.

Prospective supervisor: Pavel Pecina <pecina@ufal.mff.cuni.cz>