



Ústav formální a aplikované lingvistiky

Bachelor's Thesis Topics Proposals






at UFAL

Wednesday 27 February 2019,
15:40, S3



ÚFAL overview

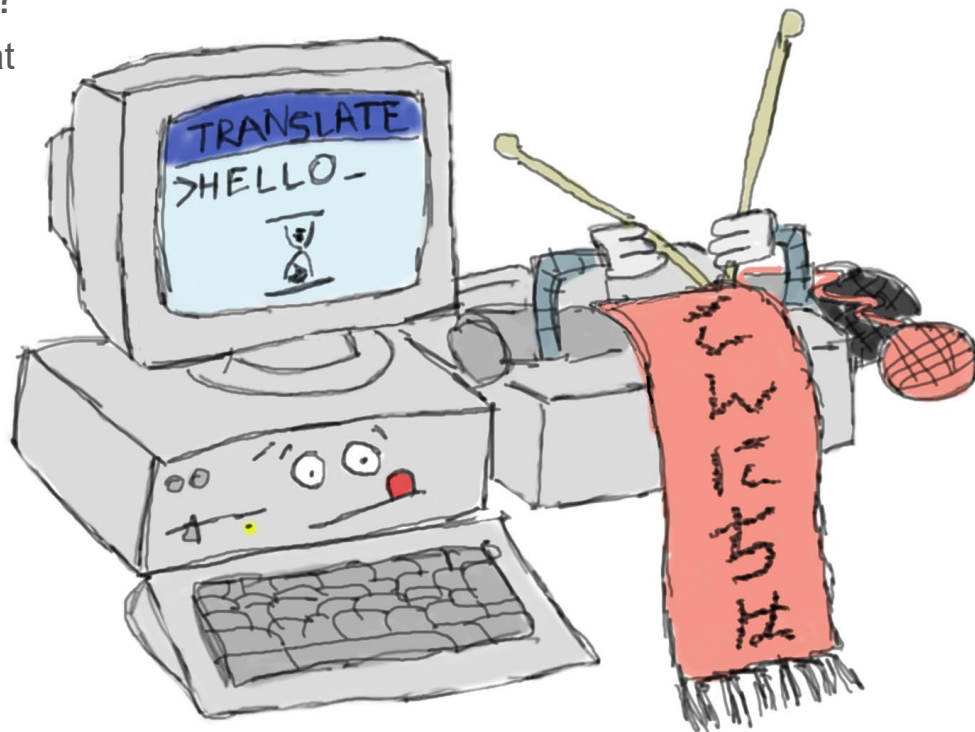


- choose your mix of:
 - Natural language processing (including Big Data)
 - Deep learning
 - Computational linguistics
 - Dialogue systems
- amazing concentration of [experts](#), shared task winners (CoNLL, WMT,...)
- many international projects and industrial cooperation
- internship support (   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



Strojový překlad (Neural Machine Translation, NMT)

- V anglicko-českém překladu jsme [nejlepší](#) na světě ([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
- Dvojazyčný Google Doc
 - textový editor pro dvousloupcové dokumenty (smlouvy ap.) s NMT
- Využití NMT pro [generování poezie](#) nebo překlad do ostravštiny/hantecu



Hraní si s jazykem

- generovat kartičky do hry iKnow (otázka, 3 nápovědy, řešení)



- automaticky vytvářet nebo řešit jazykové testy (FCE) nebo matematické slovní úlohy
- generovat textový popis ze strukturovaných údajů (např. předpověď počasí, popis trasy)
- generovat šifry či hádanky (např. synonyma, dvojznačná slova...)
- generovat diskuzní příspěvky "na iDnes" (rasistický, pro/protizemanovský...)
- spojení receptů (sloučit 3 podobné recepty na knedlík do jednoho)
- generovat jednoduchou křížovku (hřebenovka) včetně legendy
- převod textu minulost/budoucnost, vykání/tykání...
- ...a další na <http://ufal.cz/rudolf-rosa/projekty>

Vyhledávání nových slov v textech

- V nejrůznějších médiích se denně objevují nová slova.
- Chceme je mít všechny ve slovníku? Většinu ano.
- Dosud se to (někde) dělá ručně!
- Možnost spolupráce na tvorbě opravdových slovníků.

- **přepodstatňovatel**
- **shelfie** (ne, to není překlep!)
- **haldoviště**
- **trojchřtání**
- **travoltácky**
- **biozklamání**

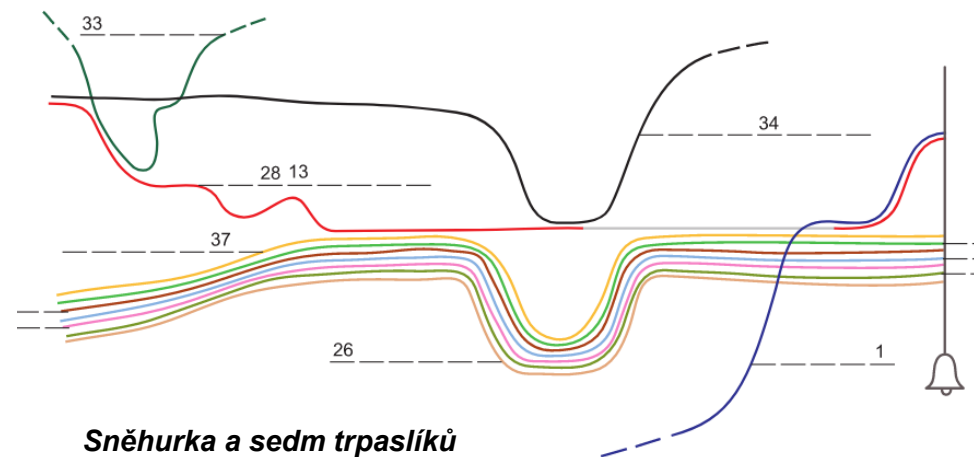
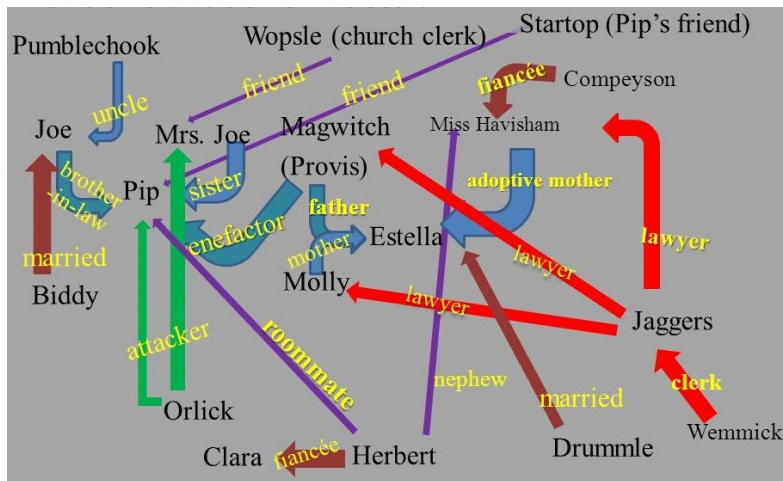
- Víte, že v každých novinách je v průměru asi 4% slov, která neznají současné slovníky?



Photo: [MSUB Library](#)

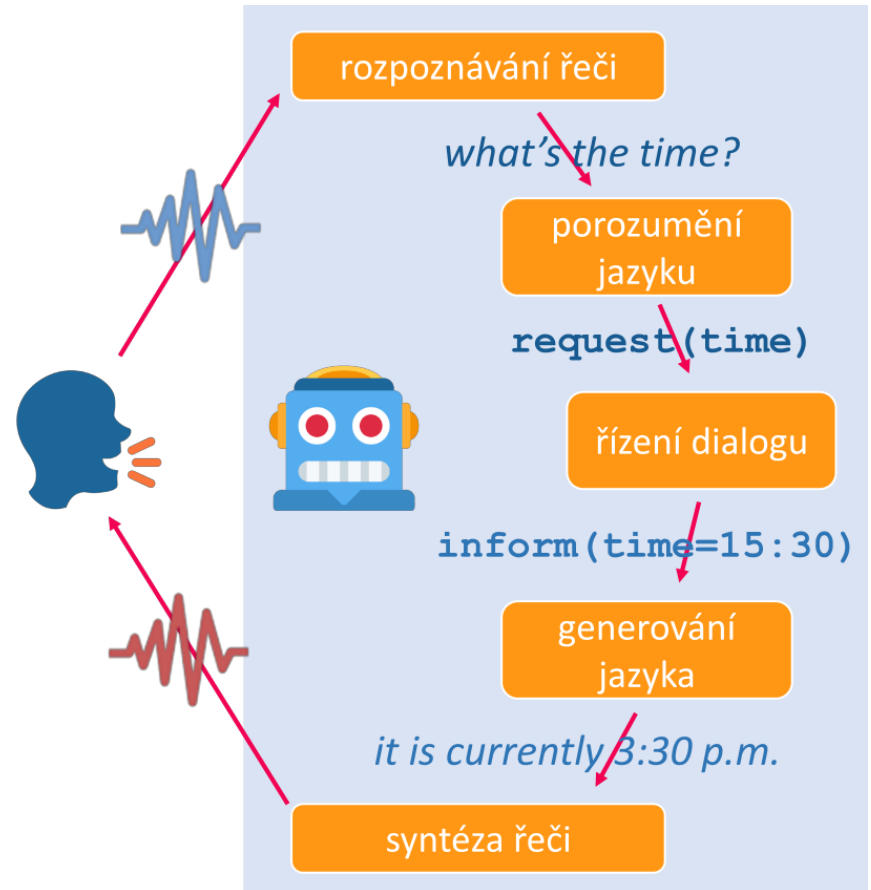
Extrakce informací z knihy

- vytvoření aplikace, která na základě vloženého textu/knihy vypíše základní informace o postavách a ději
- spousta možností od triviálních (nalezení hlavních postav) až po velmi složité a téměř nemožné (určení kdy a kde se která osoba nachází)
 - hlavní postavy
 - vztahy mezi postavami
 - časová osa: kde se o kom mluví, jak se jednotlivé postavy spolu setkávají
 - časová osa: kde se odehrává aktuální děj?



Dialogové systémy – mluvíte s počítačem

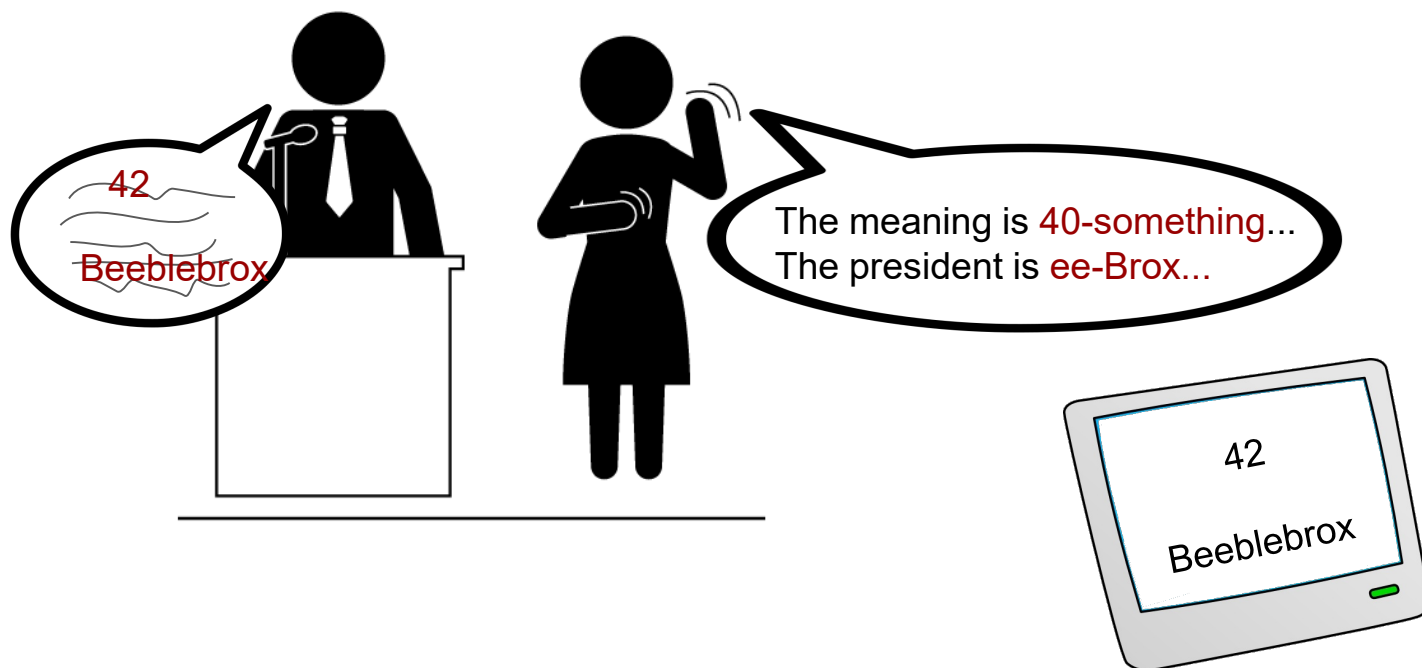
- Naučte počítač mluvit o konkrétním tématu:
 - počasí
 - městská doprava
 - televizní program
 - filmy
 - hudba
 - zprávy
 - ...
- Naučte počítač zdvořilostní konverzaci
 - oblafněte Turingův test
- Nechte počítač, ať se to naučí sám
 - výběr vhodné odpovědi z velké databáze
 - strojové učení, neuronové sítě



Více o dialogových systémech
na předmětu <http://ufal.cz/npfl123> !

Našeptávač pro tlumočníky

- Tlumočníci mají potíže, když někdo řekne mnoho čísel, vlastních jmen, ap.



- Cíl: Vytvořit specializované rozpoznávání řeči na pojmenované entity, tlumočnickovi ukazovat nápovědu: konkrétní věci, které právě zazněly.
- Forma: Jednodušší: Server-based. Užitečnější: Off-line mobile app.
 - Může využít i podklady (slidy, materiály) dodané předem.

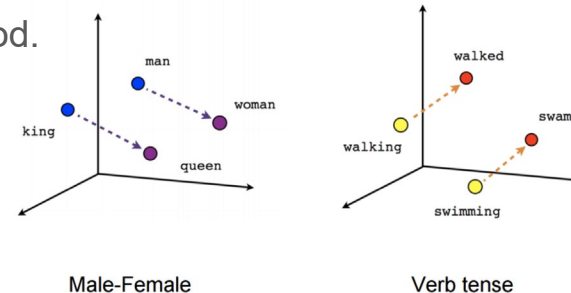
Pestrost jazyka, modelování prostoru vět

- 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.

- Nepatrnou úpravou lze význam věty **zásadně změnit**.
 - Včera **nep**řišel Novák pozdě. **E**. Novák má další pozdní příchod.

- Neuronové sítě umějí větu převést do mnohazměrného spojitého prostoru.
 - Pro slova je slavný word2vec ---->

- Dají se v tomto prostoru najít směry nebo transformace, které **odpovídají “normálním lingvistickým úpravám”**?
 - Nováka kritizovali. -(drsně)-> Nováka seřvali.
 - Nováka kritizovali. -(otočit význam)-> Nováka chválili. Nováka nekritizovali.
 - Nováka kritizovali. -(neurčitě)-> Někomu tam něco vytýkali.



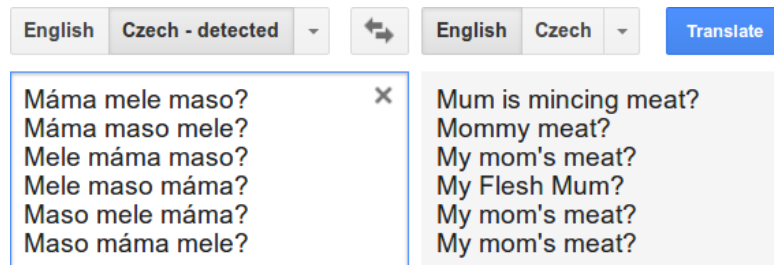
- První cíl: Za pomoci crowdsourcingu sestavit velký korpus úprav vět. (Bc)
- Druhý cíl: Promítnout sebrané věty do spojitého prostoru. (Mgr)
- Třetí cíl: Namapovat na sebe prostor úprav a spojitý prostor vět. (PhD ;-)

Strojový překlad s porozuměním obsahu

- Lidé strojovému překladu věří někdy až moc:



- Překlad se opravdu podstatně zlepšil, ale o *porozumění* nelze mluvit:

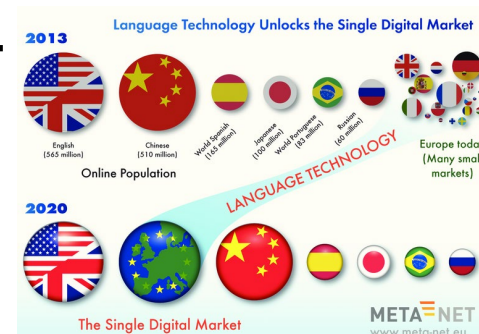


Cíl: Trénovat NMT technikou **multi-tasku**, aby zohlednilo **něco z významu**:

- Zachování sémantických rolí (kdo dělal komu co), pojmenované entity, ...
- Odpovídání na otázky o obsahu věty...

Překlad z více zdrojů; Multi-Source MT

- Cíl EU: Vše dostupné ve všech jazycích.
- Strokový překlad stále potřebuje ruční kontrolu.



Czech English Hindi Detect language ▾



English Czech Hindi ▾

Translate

How the Dutch cabinet works?
How the drinks cabinet works?
How the green cabinet works?



87/5000

Jak funguje holandská vláda?
Jak funguje nápojová skříň?
Jak funguje zelená skříň?



- **Zjednoznačnění** provedené při revizi překladu do češtiny může pomoci strojovému překladu do dalších jazyků.

Cíl: Navrhnout architekturu NMT, která umí zpracovat **libovolnou podmnožinu jazykových mutací** dané věty.



Master Thesis Topics Proposals

at UFAL

Wednesday 27 February 2019,
15:40, S3

Spoken Language Translation, Summarization

- **Neural machine translation** has made a big leap in translation quality.
 - This was thanks to end-to-end training, removal of independence assumptions.
- NN methods in **automatic speech recognition (ASR)**.

Topic 1: Train ASR+MT end-to-end.

Challenge: Make use of (independent) MT and speech data. Network architecture.

- Online meetings (skype, hangouts, gotomeeting) are very popular.
- **Minutes** are needed to keep track of what has been agreed upon.
- Writing minutes makes people busy during the call and is tedious afterwards.

Topic 2: Design (NN-based) **speech summarization** to create compact minutes.
Automatically populate pre-defined meeting agenda.

Visualization of deep learning components

A good visualization of DL-based models can lead to intuitions which in turn help us understanding what is going on under the hood of the “magic” black box.

Ideally, the student will:

1. develop a framework for visualizing different network components, and
2. use the framework to analyze a few models (e.g. NMT) on various inputs, trying to find patterns characterizing some (e.g. linguistic) features.

Neural information search

Attention models work greatly for the searching for answers for natural language questions in coherent text. However, they require a lot of training data which is available in English only. The question is what should we do if we want the model to work in another language, e.g., Czech?

The thesis would include:

- Implementing and train a model for answer span selection
- Exploring technique for transfer learning, multi-task learning

Machine Comprehension

Machine Comprehension (MC) answers natural language questions by selecting an answer span within an evidence text. The AllenNLP toolkit provides the following MC visualization, which can be used for any MC model in AllenNLP. This page demonstrates a reimplementation of BiDAF (Seo et al, 2017), or Bi-Directional Attention Flow, a widely used MC baseline that achieved state-of-the-art accuracies on the SQuAD dataset (Wikipedia sentences) in early 2017.

Enter text or

Passage

The Matrix is a 1999 science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world."

Question

Who stars in The Matrix?

Answer

Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano

Passage Context

The Matrix is a 1999 science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world."

Model internals (beta)

Generating source code from description

- Task of generating source code in Python based on its description in natural language - questions mined from StackOverflow
- International challenge announced by CMU in Pittsburgh this summer, released a large dataset

<https://conala-corpus.github.io>

```
{  
  "intent": "How do I check if all elements in a list are the same?",  
  "snippet": "len(set(mylist)) == 1",  
  "question_id": 22240602  
}
```



2 topics related to **Modelling Language Evolution**

- **1. Word migration across languages**

- Words frequently borrowed from one language to another
- Evolution of natural language vocabulary can hardly be captured by a human, because
 - zillions of morphemes in thousands of languages
 - all kinds of sound changes along the way, different alphabets...
- No, we don't want just to reproduce the work of etymologists for individual languages. We want to find a ML model that fits the evolution of the natural language as a world-wide phenomenon.

- **2. Competition in Natural Languages**

- Different languages are just different solutions to basically identical communication needs.
- Like in living nature, possible solutions of particular “tasks” compete with each other, which results in a dynamic equilibrium. We want to model the balancing mechanisms mathematically.
- Examples:
 - Typological view: fixed word order vs. richer morphology
 - Word-formation view: different suffixes competing with each other

- **In both cases:**

- Timely topics: massively multilingual data available only in the last few years
- Well-located too: 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
- If interested, let's meet at **11:00 Wednesday May 23, room 409**

Deep Understanding by Deep Learning

Despite applications like Machine Translation, true language understanding is still elusive. The focus of the thesis will be to analyze plain text in Czech and English to a knowledge representation graph by using supervised training with DNNs. Basic tools are available (up to natural language syntax), but semantic and knowledge extraction part is unsolved and will be the main problem to tackle. Datasets are available for at least two meaning/knowledge graph types (trees/DAGs).

Wikification

Wikification is the process of annotating the mentions of concepts in a document with the URL of the Wikipedia page about that concept. Such a process can be used to improve annotation of existing Wikipedia articles (even cross-lingual) or for annotating any user text. Training material is large - the entire Wikipedia!

Foucaultovo kyvadlo

Tento článek je o fyzikálním experimentu. O knize Umberta Eca pojednává článek *Foucaultovo kyvadlo (kniha)*.

Foucaultovo kyvadlo, pojmenované po francouzském fyzikovi J. B. Léonu Foucaultovi, představuje důležitý experiment potvrzující otáčení planety Země kolem své osy.

Obsah [zobrazit](#)

Historie experimentu [\[editovat | editovat zdroj \]](#)

Původní pokus byl proveden v roce 1851 v pařížském Pantheónu, kde bylo v kupoli zavěšeno závaží o hmotnosti 28 kilogramů na 68 metrů dlouhém laně. Doba kmitu kyvadla byla 16 sekund. Na závaží kyvadla byl hrot, kterým se do písku na podlaze zakresloval pohyb kyvadla. Pozorovatelé tak mohli vidět, jak se postupně mění rovina kyvu.

V roce 1851 bylo všeobecně známo, že se Země otáčí. Bylo také pozorováno zploštění Země na pólech. Foucaultovo kyvadlo však bylo prvním jasně viditelným důkazem a způsobilo tak velkou senzaci jak v odborných kruzích, tak u široké veřejnosti.

Na obou pólech Země se rovina kyvu nemění vzhledem k okolním hvězdám, zatímco Země se jednou za den zcela otočí. Vzhledem k Zemi se tedy rovina kyvu na severním nebo jižním pólu jednou za den zcela otočí po směru nebo proti směru hodinových ručiček. Pokud bychom Foucaultovo kyvadlo zavěšili na rovníku, rovina kyvu zůstane vzhledem k Zemi nezměněna. V ostatních zeměpisných šířkách se rovina kyvu vzhledem k Zemi sice otáčí, ale pomaleji, než na pólech.



Foucaultovo kyvadlo

Tento článek je o fyzikálním experimentu. O knize Umberta Eca pojednává článek *Foucaultovo kyvadlo (kniha)*.

Foucaultovo kyvadlo, pojmenované po francouzském fyzikovi J. B. Léonu Foucaultovi, představuje důležitý experiment potvrzující otáčení planety Země kolem své osy.

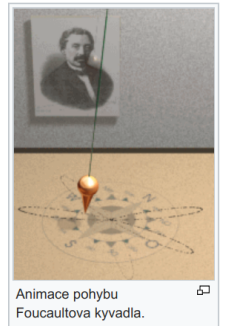
Obsah [zobrazit](#)

Historie experimentu [\[editovat | editovat zdroj \]](#)

Původní pokus byl proveden v roce 1851 v pařížském Pantheónu, kde bylo v kupoli zavěšeno závaží o hmotnosti 28 kilogramů na 68 metrů dlouhém laně. Doba kmitu kyvadla byla 16 sekund. Na závaží kyvadla byl hrot, kterým se do písku na podlaze zakresloval pohyb kyvadla. Pozorovatelé tak mohli vidět, jak se postupně mění rovina kyvu.

V roce 1851 bylo všeobecně známo, že se Země otáčí. Bylo také pozorováno zploštění Země na pólech. Foucaultovo kyvadlo však bylo prvním jasně viditelným důkazem a způsobilo tak velkou senzaci jak v odborných kruzích, tak u široké veřejnosti.

Na obou pólech Země se rovina kyvu nemění vzhledem k okolním hvězdám, zatímco Země se jednou za den zcela otočí. Vzhledem k Zemi se tedy rovina kyvu na severním nebo jižním pólu jednou za den zcela otočí po směru nebo proti směru hodinových ručiček. Pokud bychom Foucaultovo kyvadlo zavěšili na rovníku, rovina kyvu zůstane vzhledem k Zemi nezměněna. V ostatních zeměpisných šířkách se rovina kyvu vzhledem k Zemi sice otáčí, ale pomaleji, než na pólech.



Text segmentation

Text segmentation is the process of dividing written text into **meaningful units**, such as **words**, **sentences**, or **topics**. The term applies both to mental processes used by humans when reading text, and to artificial processes implemented in computers, which are the subject of natural language processing.

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 podlehy **milióny** lidí. Latinský název chřipky – influenza (v angličtině obvykle zkracováno na flu) – pochází z italského a původně šlo o termín označující viru v nepříznivé astrologické vlivy, **influentiae**, coby příčiny nemoci. 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 imunitizaci 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 dále 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 viru typu A se kromě vysoké mutagenity vyskytuje i nebezpečná možnost rekombinace: pokud dva různé subtypy viru napadnou tužeb 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 ptáci chřipka šíří nejnáze, a také rozsáhlého chovu prasat na jednom území – prasata jsou infikovatelná jak savci, tak i většinou typů ptáčích chřipek (i těch, co většinu savců nenapadají), což zvyšuje pravděpodobnost nových, „radikálních“ konstrukcí viru, které by mohly být nebezpečné člověku. Často se při klasifikaci nemoci odlišují viry tzv. ptáčích chřipky – která napadá hlavně ptáky a savce jen omezeně, resp. téměř vůbec – od chřipky napadajících savce. Je zde ovšem vždy riziko mutace, které udělá z ptáčích 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 podlehy **milióny** lidí.

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

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 imunitizaci 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 dále 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.

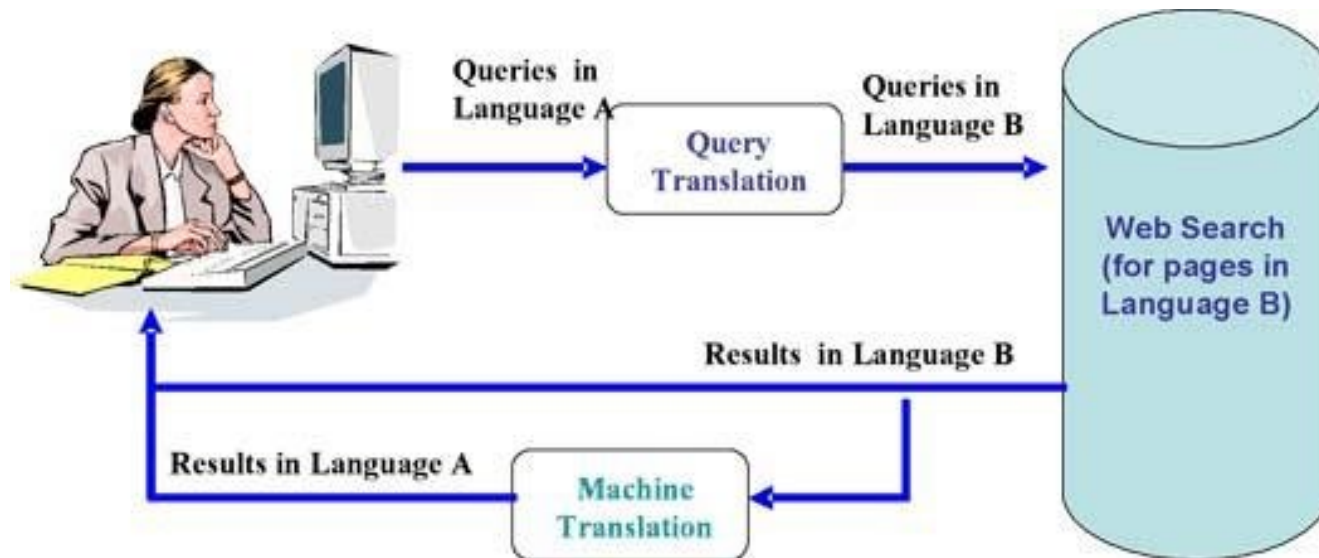
Existují ještě další variace viru, a proto jsou specifické chřipkové kmenové oddíly identifikovány standardním názvoslovím specifikujícím typ viru, geografickou polohu prvního výskytu viru, rok izolování, pořadové číslo izolování a subtypy HA a NA (např. názvy „A/Moscow/10/99 (H3N2)” či „B/Hongkong/330/2001”).

Variabilita a rekombinace

U viru typu A se kromě vysoké mutagenity vyskytuje i nebezpečná možnost rekombinace: pokud dva různé subtypy viru napadnou tužeb 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 ptáci chřipka šíří nejnáze, a také rozsáhlého chovu **prasat** na jednom území – prasata jsou infikovatelná jak savci, tak i většinou typů ptáčích chřipek (i těch, co většinu savců nenapadají), což zvyšuje pravděpodobnost nových, „radikálních“ konstrukcí viru, které by mohly být nebezpečné člověku. Často se při klasifikaci nemoci odlišují viry tzv. **ptáčích chřipky** – která napadá hlavně ptáky a savce jen omezeně, resp. téměř vůbec – od chřipky napadajících savce. Je zde ovšem vždy riziko mutace, které udělá z ptáčích chřipky chřipku napadající i savce a člověka.】

Multilingual embeddings for cross-language search

Cross-language information retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a language (B) different from the language of the user's query (A).



The traditional approach to CLIR is translation of queries (in A) into the language of documents (B). We would like you to solve the same task by using multilingual (word embeddings).

Neural information retrieval

Deep Learning has not affected the field of information retrieval (i.e. web search) much yet but there is a plenty of potential applications of neural technologies in information retrieval:

- Learning query/document representations (embeddings)
- Matching query/document representations
- Ranking search results
- Summary/snippet construction

