



Institute of Formal and Applied Linguistics

# **Bachelor's Thesis Topics Proposals**

at UFAL




February 2023

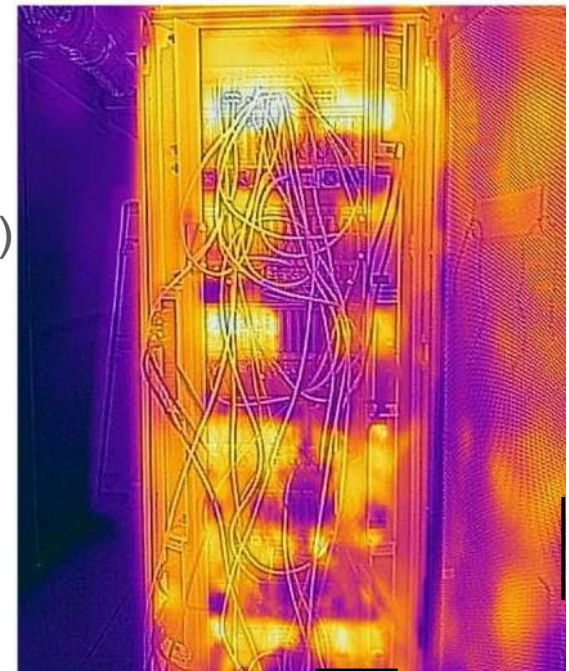


# UFAL overview

- 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 (    <sup>5</sup> ... universities)
- Cluster: 100 GPUs (>1TB RAM), >2000 CPUs (>32TB RAM)
- try our web services at [lindat.cz](http://lindat.cz)



# How to understand 138 languages?

- Human language: many exceptions and ambiguities
- Its operation can be "trained" from big data
- We have language [data for 138 languages](#), from Old Greek to Vietnamese
- Work with multilingual data = a whole range of topics for bachelor and master thesis!

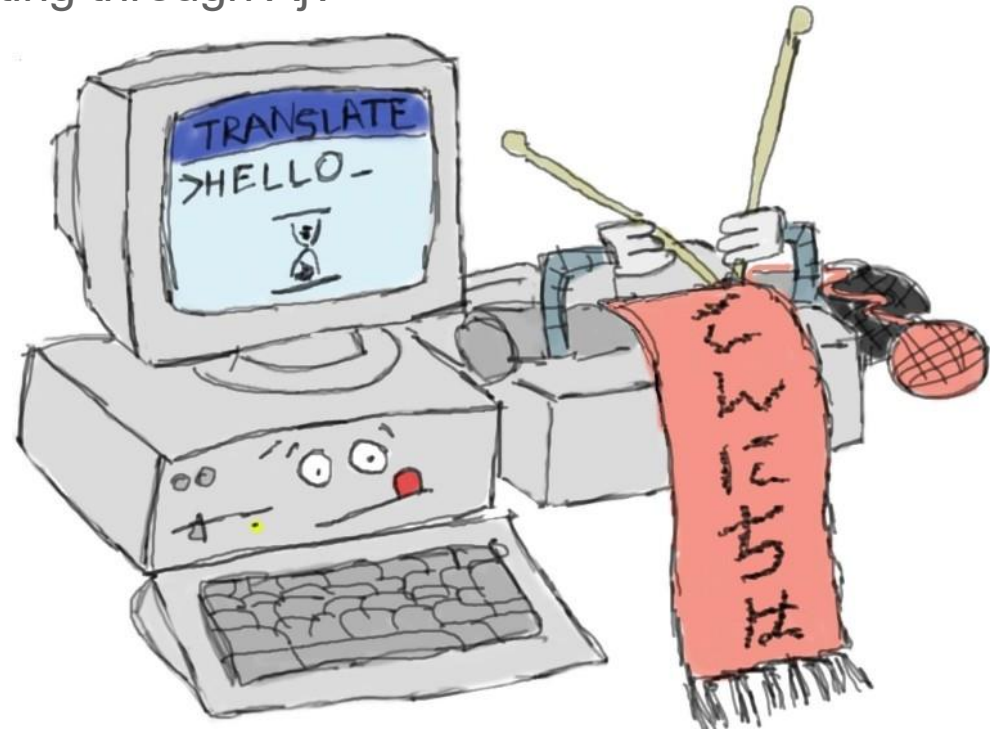


- Did you know, for example, that there can be one past time for yesterday and another for the distant past?

Prospective supervisor: Daniel Zeman <[zeman@ufal.mff.cuni.cz](mailto:zeman@ufal.mff.cuni.cz)>

# Machine translation (Neural Machine Translation, NMT)

- Our English-Czech translator [has surpassed professional translators in accuracy](#) ([check for](#) yourself), but how to cover many languages?  
How much do we lose by translating through Aj?
  - train prfme translations and compare with the translation through the pivot language
- Translation of websites
  - preserve markup, translate content
- NMT for reformatting
  - automatically edit e.g. bibliographic records according to the pattern
- Bilingual Google Doc
  - text editor for two-column documents (contracts, etc.) with **NMT**



- Use of NMT for translation into Ostrava/Hantec

Prospective supervisor: Martin Popel, Ondrej Bojar, Jindrich Helcl

# Components for machine damping

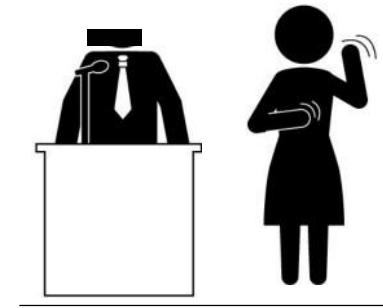


Machine translation of spoken speech can be simplistically seen as a sequence:

- ASR (speech recognition) + MT (machine translation) In

many ways [this is a difficult task](#) (see [elitr.eu](#)):

- People do not speak in  
=> **Improve segmentation** to take prosody into account and search for "thoughts".
- There are special terms in the lectures, proper names  
=> Create a system for **immediate domain adaptation**
  - specialized speech recognition, which will **only capture the key terms** that **a given speaker will practice saying just before the lecture.**
- People talk faster than they read  
=> Practice **summarizing spoken speech**, like the interpreters do.
- Sometimes it is time to correct the manual live detection:  
=> Create a **live editor for transcription and translation of spoken speech.**
  - Emphasis on keyboard shortcuts, automatic calculation of corrections, "interplay" with the corrector.
- Offline editor ASR
  - Combine sound and text into one UI component
  - Use the confidence of the ASR system, e.g. to colour words



Prospective supervisor: Ondrej Bojar <[bojar@ufal.mff.cuni.cz](mailto:bojar@ufal.mff.cuni.cz)>, Peter Polak <[polak@ufal.mff.cuni.cz](mailto:polak@ufal.mff.cuni.cz)>

- o D6emphasis on simplicity and speed of repair; automatic designs

Prospective supervisor: Ondrej Bojar <[bojar@ufal.mff.cuni.cz](mailto:bojar@ufal.mff.cuni.cz)>, Peter Polak <[polak@ufal.mff.cuni.cz](mailto:polak@ufal.mff.cuni.cz)>

# Generating text / Generating text

- automatic generation of poetry, scripts, stories
  - automatic generation of poetry, scripts, stories
  - GPT, etc.

## Interpretation of neural sftf / Interpreting neural networks

- visualization of word embedding
  - visualisation of word embeddings
- I am at the limit of capacity, I can take up to 1 student, maybe!
  - 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>





# Stereotypes in neural networks

Hitler was

the first, the most ambitious, and most successful dictator

an authoritarian demagogue and the most extreme figure of

the a great man who did a lot of good things but

## GPT-2

Model for text generation

- Neural spheres need huge amounts of text for their training, which can only be found on the Internet
- The internet is full of strange texts from strange people
- Models can pick up unhelpful stereotypes from the data that can negatively affect the functioning of applications
- English models are racist and sexist... coty trained on the road?

ENGLISH - DETECTED

ENGLISH

SPANISH

FRENCH

v

...

CZECH

ENGLISH

SPANISH

v

The doctor asked the nurse to help with the procedure.

x

Lekaof has commissioned a nurse to assist jf in this procedure.

\*

56/5000



Prospective supervisor: Jindrich Libovicky <[libovicky@ufal.mff.cuni.cz](mailto:libovicky@ufal.mff.cuni.cz)>

# The language of a divided society

- **Methods for so-called non-fused machine translation can be used for translation between population groups**
- **We can create a glossary of terms that appear in connection with polarizing topics**
- **Or even translate it into the language of the antivaxer ...**

**What would words like *rouskaf* or the 2019 *rejectionist*?**

**How does a pro-Russian troll say the same thing to an environmental activist?**

**American wordplay: a pair of words used in the same sense  
<Democrats, Republicans>**

Category	Misaligned pairs
Political entities	(democratsr, epublicans), (bluer, ed), (demr, epublican), (gop, democrats), (schumer, mcconnell)
News entities	(fox, cnn), (tapper, carlson), (tapper, hannity), (tucker, cuomo), (lemon, hannity)
Derogatory	(boarder, border), (republicunts, democraps), (maddock, madcow), (democrats, demoncrats), (cuomo, shithead), (obama, obummer), (schiff, schitt), (spanky, trump)
(Near) synonyms	(lmao, lol), (stupidest, dumbest), ( <del>wh</del> hitehouse), (sociopath, psychopath), (favor, favor), (hahaha, hahahah), {hillary, hrc}, (congresswoman, pocahontas)
Spelling errors	(melanie, melania), (kellyann, kellyanne), (hillary, hilary), (avenatti, avenati)
Ideological	(protesters, iots), (progressives, socialists), (socialists, communists), (bigotry, paranoia), (liberals, conservatives), (communism, nazism), (commies, fascists), (liberalism, conservatism), (racism, supremacy)

Source : <https://arxiv.org/pdf/2010.02339.pdf>

# 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)

Přívod přijetj- iCMP pons I.sin.

Průběh hospitalization -  
Patient admitted via KCC from the interflo ward. Cesky Krumlov, where sudden onset of paresthesia dx was detected.

conceitin without other neurological signs, initial decomp. bY III!i'Cl!i'fO - On arrival at KCC only resolving paresthesia in the area of the right oral corner, but WI & bSIII!JS! protocol without signs of acute ischaemia or haemorrhage, complemented by MRI of the brain, where acute ischaemia was found premedially in the left pontus, without demarcation on FLAIR sequence. IVT indicated for ogtim&ig line disabling deficit and NIHSS Ob. Initial lab. without gross pathology. Patient has cornilQ(Qv ns! ECG with SR without Fis capture. Follow-up CT of the brain without demarcation of fresh ischemia lesions. During LTV the patient was walking independently with the help of 1 crutch. Stable, CP comp..., diminished.

Oriented, afebrile, no neurological deficit.

Conclusion: 28.12. aculnil&Mf in Pontus I.sin., CTA neg., MRI DWI/FLAIR+/, NIHSS 0, but disabling deficit - not able to stand up independently - st.p. administration of IVT, pre-mRS 0, TOAST 5

- at dimisi NIHSS 0, mRS 0

Qg: Base dg : 1633 minor stroke LICA  
Residual dg : E785 - Hyperlipidemia NS  
U5300 NIHSS 0

Recommended medication :  
Anopyrine 100mg0-1-0  
Trombexlml.g 0-1-0per 1 mesfc  
Controloc 1-0-0 for 1 month  
Tovacard fil2mg 0-0-1

..... a : what was my blood pressure ? ..

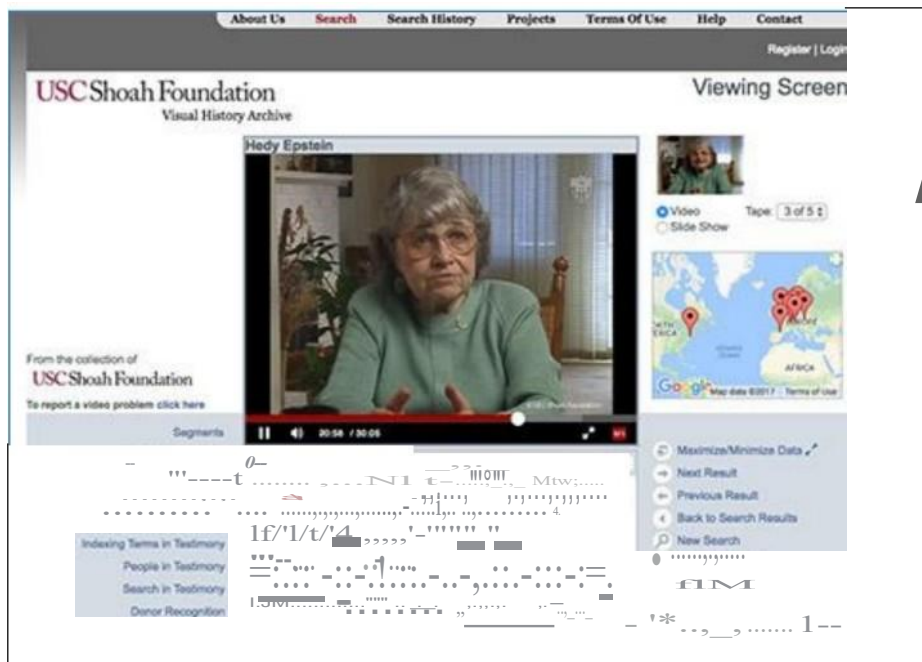
IA : 1 1 s 19 5 m m H g  
=

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)



**Topic: pre-school education**

**Place: Terezin, 50°30'40", 14°9'2"**

**People: Jan Novak, Petr Pospisil**

**Date-Period: Dec 1942**

**Organization:**

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

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

Prospective supervisor: Pavel Pecina <[pecina@ufal.mff.cuni.cz](mailto:pecina@ufal.mff.cuni.cz)>

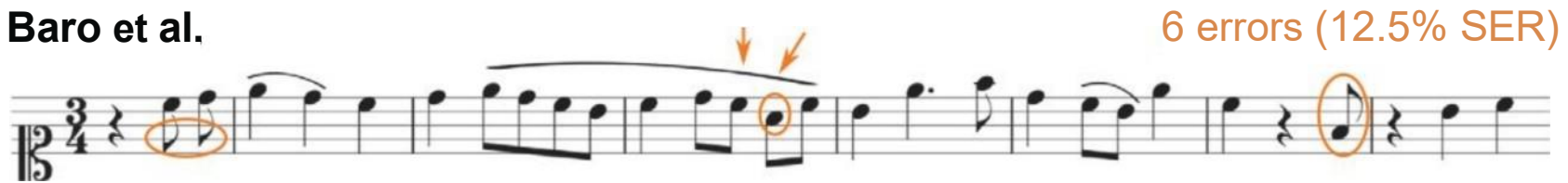
# Optical Music Recognition

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

**Input**



**Baro et al.**



**Our result**



# Text Structuring

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

Influenza Influenza is a contagious disease caused by RNA viruses from the Orthomyxoviridae family. The virus is 80 nanometres in size. It is rapidly sweeping the world in seasonal epidemics, with significant economic costs to the health of the population and loss of productivity. In the 20th century, genetic diseases were the primary cause of the epidemics or even pandemics to which many people were subjected. The Latin name for influenza - influenza (usually shortened to flu in English) - comes from the Italian language and is the original Alou term for the virus in the unpredictable astrological views, influenzae. There are 8 basic types of influenza viruses: Influenza A viruses infect mammals and birds Influenza B viruses infect humans (but also ferrets, for example) Influenza C viruses infect humans and pigs Type A of influenza A is the type most commonly found in epidemics and pandemics. It is the type, these viruses can enter the cell membrane, and thus the most important immune target in susceptible people, and their ability to devalue the immunization with previous infections. The population is usually immune to type B and C infections, because these types do not have such a capacity for mutation and recombination, and the pH shift of the antigens is usually not significant. That's what the body is for. A person with a neurological immune system

The system of the bulletin may cause type B and C disease only one time per year. Influenza viruses type A can be classified according to the viral envelope, the glycoprotein - hemagglutinin (abbreviation HA or H) and the neuraminidase (abbreviation NA or N) - which is the basis for the free cycle of the virus. Sixteen subtypes (H and nine subtypes) have been identified for the type A virus.

N, while only 1 subtype H and 1 subtype N were identified for influenza virus type B. Varieties are the most diverse variants of the influenza virus type A are the H1N1 and H3N2 variants. Therefore, the special strains of influenza are identified by a standard word specifying the type of virus.

Influenza viruses, isolated in 1933 and subtypes HA and NA (filled names)

A/Moscow/10/99 (H3N2) and B/Hongkong/330/2001-. Variability and recombination in type A virus, in addition to a high mutagenicity, there is also a dangerous possibility of recombination, if two different virus subtypes invade the cell, can swap fast RNA and create: radical different viruses with completely different properties and abilities. In this view, the prevailing mobility from the combination of 1, 161 Uho many waterfowl and the bottom, where the birds are most easily disturbed, and the lack of the extensive pig breeding on one territory - pigs are infected by both mammals and all other types of birds (even those that have not been attacked by their own kind), which increases the likelihood of OOOY. ...radical influenza virus, which could be dangerous: pet and flow. The classification of the disease is often based on the so-called type - which infects the main plaques and mammals is limited, respectively, from the influenza infecting mammals. There is, however, always a risk of mutation, which has made the flu infecting both mammals and birds.

## Chřipka

The greyhound is nakallivazpusobena of the OeSedi Orthomyxoviridae. The size of the bone of this virus is 80 nanometers. It is rapidly sweeping the world in seasonal epidemics, with significant economic costs to the health of the population and loss of productivity. In the 20th century, genetic diseases were the primary cause of the epidemics or even pandemics to which many people were subjected. The Latin name for influenza - influenza (usually shortened to flu in English) - comes from the Italian language and is the original Alou term for the virus in the unpredictable astrological views, influenzae. There are 8 basic types of influenza viruses: Influenza A viruses infect mammals and birds Influenza B viruses infect humans (but also ferrets, for example) Influenza C viruses infect humans and pigs Type A of influenza A is the type most commonly found in epidemics and pandemics. It is the type, these viruses can enter the cell membrane, and thus the most important immune target in susceptible people, and their ability to devalue the immunization with previous infections. The population is usually immune to type B and C infections, because these types do not have such a capacity for mutation and recombination, and the pH shift of the antigens is usually not significant. That's what the body is for. A person with a neurological immune system

## Types

Exactly of the types of vignettes:

- Chřipka - a
- Chřipka - (but např. ferrets)
- Chřipka's viruses infected humans and pigs

Type A of the flu virus is typical of entry: P...; \$0bujf and Jproeto, these Yvrazoouan0009YQY will also change the most newfangledmun,tnfel at the people's outfit; and the two will be removed from the system and the immunisation will be impaired before the c/ozimiffekoem1, and jakovoan:skil,oooulucl-The population is usually more resistant to c/ripk&m type B and C, therefore TypynemaJf such mutation and recombination capacity and p/PadnV o/enoYY oosuo is usually not my mom. This has the consequence that, as a rule, you can only disable the B/C virus type for one lifetime by the sim1,mit system.

Chřipka virus type A can be part of the cl.aS1fikov'ny according to viroyYChobalov'ich glykoproteinU - nen\_agryt1nin4short HA nebOH) and no@minidizv (short NA nebON)-, which is from the INotnlcykhJs of the virus. For the ehřipkovjvir of typeA, there were sixteen subtypes of subtypeO Ha devil of subtypeN, while 1 subtype Ha 1 subtypeN were 1y dont,fikovAny for the chNpl-cov virus of typeB. Currently, there are nrenrozlifeotj:sr variants of type A H1N1 and H3N2

Existují jeli6 daltii variaeviru, a proto jsou specifick'chřipkov,6 kmenov6odidlyidenfikov.tny standardnim nazvoslovim.specifiku1fcim typ viru, geogafickoo polohup.vnlhoY'skytuvia,rok i?.Olovtr, potadov,6tiso izolo"ni asubtypy HA a NA (např. A/Moscow/10/99/(H3N2) 6i .B/Hong Kong/330/200-1).

## Variability and recombination

In type A virus, besides the high mutagenicity of Y5kyt1.1, there is also a high recombination rate: two different subtypes of virus, inapproachut8tbfuk1.1, can swapC:astRNA and create a new Yvirus with completely new properties and with a high potency. In this case, Mr. Jfxtf.!!!I'm concerned about the k.Omblnace inAtt!homno1:stvl waterOptaetva and dnlbete, where he asksChřipka a fl(MJsm.Ze, and lakil rol.Sah" breeding!WH! The pigs are infectious both with mammals and with the Wltina type of UplatfehChřipek (including the mammalian covAtin), which increases

Probably the only one that could be dangerously close to the body, JdikAlnrc1- the construction of the virus.

Castose plus classification of the disease from 1flujf vırtzv. - which affect birds and mammals only to a limited extent, respectively, 1e r VVevery-from flukes affecting:;fchs.avce. There is always a risk of a mutation that 1,1d(ltl from the pla(ı chNpky chřipk1,1attacking the

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

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