

This document is part of the Research and Innovation Action “European Live Translator (ELITR)”.
This project has received funding from the European Union’s Horizon 2020 Research and
Innovation Programme under Grant Agreement No 825460.



Deliverable D7.2

Report on NLP Technologies Workshop at EUROSAT Congress

Rudolf Rosa (CUNI), Ondřej Bojar (CUNI), Dominik Macháček (CUNI),
Vilém Zouhar (CUNI), Tomáš Musil (CUNI), Michal Auersperger (CUNI)

Dissemination Level: Public

Final (Version 1.0), 30th June, 2020





Grant agreement no.	825460
Project acronym	ELITR
Project full title	European Live Translator
Type of action	Research and Innovation Action
Coordinator	Doc. RNDr. Ondřej Bojar, PhD. (CUNI)
Start date, duration	1 st January, 2019, 36 months
Dissemination level	Public
Contractual date of delivery	30 th June, 2020
Actual date of delivery	30 th June, 2020
Deliverable number	D7.2
Deliverable title	Report on NLP Technologies Workshop at EUROSAT Congress
Type	Report
Status and version	Final (Version 1.0)
Number of pages	25
Contributing partners	CUNI
WP leader	AV
Author(s)	Rudolf Rosa (CUNI), Ondřej Bojar (CUNI), Dominik Macháček (CUNI), Vilém Zouhar (CUNI), Tomáš Musil (CUNI), Michal Auersperger (CUNI)
EC project officer	Alexandru Ceausu
The partners in ELITR are:	<ul style="list-style-type: none"> ▪ Univerzita Karlova (CUNI), Czech Republic ▪ University of Edinburgh (UEDIN), United Kingdom ▪ Karlsruher Institut für Technologie (KIT), Germany ▪ PerVoice SPA (PV), Italy ▪ alfatraining Bildungszentrum GmbH (AV), Germany
Partially-participating party	<ul style="list-style-type: none"> ▪ Nejvyšší kontrolní úřad (SAO), Czech Republic

For copies of reports, updates on project activities and other ELITR-related information, contact:

Doc. RNDr. Ondřej Bojar, PhD., ÚFAL MFF UK bojar@ufal.mff.cuni.cz
Malostranské náměstí 25 Phone: +420 951 554 276
118 00 Praha, Czech Republic Fax: +420 257 223 293

Copies of reports and other material can also be accessed via the project's homepage:

<http://www.elitr.eu/>

© 2020, The Individual Authors

No part of this document may be reproduced or transmitted in any form, or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the copyright owner.



Contents

1	Executive Summary	4
2	Workshop Description	5
2.1	Target group of the Workshop	5
2.2	Goal of the Workshop	5
3	Workshop Structure	6
3.1	Presentation: Linguistics and Language Processing	6
3.2	Demo: Named Entity Recognition	6
3.3	Demo: Question Answering	7
3.4	Demo: Machine Translation	7
3.5	Demo: Cross-Lingual Information Retrieval	7
3.6	Demo: Speech Recognition & Speech Translation	7
3.7	Discussion: Use language tools at your SAI!	8
3.8	Wrap up	8
4	Workshop Materials	9
4.1	Workshop website	9
4.2	Workshop worksheets	12
4.3	Workshop slides	12
5	Conclusion	25



1 Executive Summary

This deliverable reports on the preparation of the LangTools workshop at EUROSAT Congress 2020, aimed at presenting NLP Technologies to supreme audit institution (SAI) representatives.

As the congress was shifted to 2021 due to Covid-19, the workshop, planned for June 2020, has not yet taken place. Nevertheless, all materials, demos and presentations have already been prepared, are described in this deliverable, and some of them are directly attached.

The workshop consists of a frontal presentation on NLP technologies, five interactive demo sessions, and a mediated discussion.

The workshop has been rehearsed several times, including a mock workshop with employees of Czech SAO, i.e. in a very realistic setting considering the intended target group of the workshop.

The congress and thus the workshop is now planned for June 2021.



2 Workshop Description

2.1 Target group of the Workshop

The target group of the workshop are heads and employees of various supreme audit institutions, who would be responsible for a decision to incorporate NLP tools into the auditing practice of their institution.

The participants actively manifested their interest in current NLP technologies by voluntarily registering for the workshop. However, the participants are not expected to have any prior knowledge of NLP technologies.

The participants are not expected to be proficient computer users. They are also not expected to be fluent in English.

2.2 Goal of the Workshop

The workshop aims to familiarize the participants with the current state, quality and availability of automated natural language processing technologies, with a focus on machine translation, search in foreign-language documents, speech transcription and translation, and automated analysis of texts.

The workshop also seeks to gather ideas for areas of auditing practice which could be improved with new technologies. This could mean making them faster, broader or more accurate, e.g. by easily employing documents written in foreign languages.



3 Workshop Structure

The workshop consists of an introductory presentation, five interactive demo sessions, a moderated discussion, and a final wrap up.

The participants are equipped with pre-set laptops. The presentations are run in English, but the participants have the option to watch live English subtitles as well as live translations of the subtitles into their languages.

To assist the participants, there are also *facilitators*, i.e. NLP professionals who help and advise the participants if needed. The maximal number of participants is 30, with 6 facilitators available, i.e. there is one facilitator per 5 participants.

The whole workshop is planned for 2.5 hours, with the following programme:

Block 1	45 min
Presentation: Linguistics and Language Processing	15 min
Demo: Named Entity Recognition	15 min
Demo: Question Answering	15 min
Break	15 min
Block 2	45 min
Demo: Machine Translation	15 min
Demo: Cross-Lingual Information Retrieval	15 min
Demo: Speech Recognition & Speech Translation	15 min
Break	10 min
Block 3	30 min
Discussion: Use language tools at your SAI!	15 min
Wrap up	15 min

In further sections, we describe the contents of the individual parts of the workshop. The workshop materials are attached in Section 4.

3.1 Presentation: Linguistics and Language Processing

The main presenter welcomes the participants and introduces them to current NLP technologies, from the perspective of their potential employment in SAI auditing practice.

Various technologies are mentioned, such as mobile phone writing assistance, speech recognition, machine translation and spoken language translation.

The presenter also presents Technology Readiness Levels (TRL) and explains that various technologies are currently at various TRLs, with some already being massively used in practice while others are still being researched. In the subsequent demos, each tool is then accompanied by its approximate TRL, so that the participants get an idea whether this tool could be easily implemented immediately, or whether several months or years of research and development would be necessary prior to its implementation in practice.

The presenter then explains the structure of the rest of the workshop and gives floor to the demo sessions.

3.2 Demo: Named Entity Recognition

In the first demo session, the presenter shows several documents with named entities recognized and highlighted. He shows how highlighting named entities helps with understanding and orientation in the document.

After this first demonstration, the presenter briefly explains what named entities and named entity recognition is, and what it can be useful for.



Finally, the participants are encouraged to open the displaCy Named Entity Visualizer (<https://explosion.ai/demos/displacy-ent>), try running it on various text and examine the results. They are also asked to look for errors in the output of the tool, so that they understand both the advantages as well as limitations of automated processing of language.

3.3 Demo: Question Answering

The presenter briefly motivates the problem of question answering and opens the BERT-based Question Answering Demo (<https://zilinec.me/bert/>). He shows the participants how to use the tool with the default document and several suggested questions, for which the tool returns the correct answer as a snippet of the input document.

The participants are then led to reuse one of the documents from the previous demo to show an alternative way of finding the same information as before in the document. The final task is try out the tool on a random Wikipedia article to demonstrate that the tool really is general and can work well practically with any document.

3.4 Demo: Machine Translation

After the break, the workshop continues with a demo of machine translation. Since this is something that all the participants presumably know quite well, the presenter goes directly to demonstrations.

The demo revolves around the Ptakopët tool (<https://ptakopet.vilda.net/?p=sao>), which uses forward and backward translation to help the user to create a text in foreign language and check that the meaning is mostly correctly retained. The main task for the participants is to produce an e-mail message in Estonian with the tool, putting in the text in English, using automated translation to get the text in Estonian, checking the preservation of the correct meaning on a translation of the Estonian text back to English, and gradually modifying the source English text until the translation seems to be correct.

3.5 Demo: Cross-Lingual Information Retrieval

In the fourth demo session, the presenter demonstrates a web-based cross-lingual information retrieval tool, available at <http://bit.ly/ws-clir>.

The presenter briefly introduces the problem and tool, shows a simple search query, and then presents the participants with three tasks, in which they have to search the document collection to obtain a particular piece of information. The participants try to solve the tasks with the help of the facilitators. The presenter then concludes the session by showing how the tasks can be solved.

The tool is implemented using Apache Solr information retrieval toolkit and is adapted to the needs of SAIs. The search collection contains audit documents from various SAIs and in various languages, machine-translated to several languages to allow for cross-lingual search. The tool was developed specifically for this workshop. It is maintained in a GitHub repository (<https://github.com/ufal/clir>).

3.6 Demo: Speech Recognition & Speech Translation

The last demo is the most interactive, show-casing technologies central to the ELITR project. The whole session revolves around a task in which the participants have to communicate through the alfaview speech translation service to solve a task of describing and guessing a word, i.e. a sort of a simple game.

The participants are divided into pairs and handed a sheet of paper with a list of words. In the pair, they then connect through AlfaView, using the provided laptop and headset. Then they take turns, with one participant describing one of the words on the list and the second participant having to guess the word. To really experience the technology, the participants are encouraged to turn off the sound output for the describing participant, so that the guessing



participant can only see the automated transcript of the first participant's speech. Thus, if the pair manages to communicate and guess the word correctly, they personally experience that the used technology is already mature enough to be relied on to some extent.

3.7 Discussion: Use language tools at your SAI!

For this session, the participants are organized into several groups, preferably based on their fluency in a common language to ease discussion. Each group is accompanied by at least one facilitator who is fluent in the discussion language used in the group.

The participants are encouraged to think and discuss how language tools could help at their SAI, discuss possible uses of language tools and suggest new useful language tools. The facilitators are instructed to intervene as little as possible if the discussion flows well, but to actively lead the discussion if it does not work well or diverges from the topic too much, asking the participants specific questions and suggesting specific solutions. They also answer questions of the participants, and provide more information where appropriate.

The goal of the discussion is to come up with realistic ideas on future possibilities of development and integration of NLP tools useful for SAI practice.

3.8 Wrap up

In the last session, the main presenter stops the discussions and lets each group briefly summarize what they talked about and what ideas they came up with. He then comments on these outcomes, possibly adding some more interesting information. Finally, he briefly summarizes what the workshop was about and what the participants learned, and the workshop ends.



4 Workshop Materials

We prepared a range of materials for the workshop. The core material are workshop slides (Section 4.3), which are shown throughout the workshop and which we also enclose here. The participants also get web-based worksheets (Section 4.2).

A secondary material is the webpage of the workshop (Section 4.1). An information brochure was also planned but not yet realized.

4.1 Workshop website

The participants are invited to the workshop by a brief web presentation, which can be found at <https://www.eurosai2020.cz/workshop07.html>.

The website contains a short video clip, produced by NKÚ, show-casing one of the tools presented at the workshop, namely cross-lingual information retrieval. It briefly introduces the workshop, and then links to a Workshop Description document (<https://www.eurosai2020.cz/workshops/pdf/ws07-description.pdf>) with a one-page explanation of the motivation and aims of the workshop.

On the following two pages, we enclose a copy of the webpage and the Workshop Description document.



Workshop USING LANGUAGE IT TECHNOLOGIES IN AUDITS - invitation video



WORKSHOP 7 USING LANGUAGE IT TECHNOLOGIES IN AUDITS

MAIN ORGANIZER: Charles University of Prague (ELITR project)
In co-operation with: SAI Czech Republic
CONTACT: Ms Tereza Vojtěchová [✉vojtechova@ufal.mff.cuni.cz](mailto:vojtechova@ufal.mff.cuni.cz)
Mr Ondřej Bojar [✉bojar@ufal.mff.cuni.cz](mailto:bojar@ufal.mff.cuni.cz)

"Communicate automatically in all languages!"



Translate automatically any document!



Automatic question answering!



Ask questions in any language!



Search across languages!



Make phone calls in any language!



Present to international audience!





7 / Using language IT technologies in audits

Main organizer: Charles University of Prague (ELITR project)
Contact: Ms Tereza Vojtechova vojtechova@ufal.mff.cuni.cz
Mr Ondrej Bojar bojar@ufal.mff.cuni.cz
Co-organizers: SAI Czech Republic

Today we are experiencing a boom of computer technologies, we all have smart mobile phones, cars, and/or even whole households, but what about language and so called natural language processing? Could it be somehow made very fast and automated? Could language processing technologies be useful for audits and different fields? Could they be useful for you?

What is our aim?

This workshop will give you **an idea about the latest language technologies, their continuous improvement and usefulness in many different areas.**

- Do you need to translate large volumes of documents? It already works, and automatically.
- Do you need to search in documents written in a language you do not understand? It works and it is reliable.
- Do you need to read large numbers of documents, in search for particular values, dates or facts (so-called text analytics)? It works and it is customizable.
- And mainly, do you want to take part in discussions which are held in a language you do not speak? It works and you will see a demo.

During our workshop you can try out existing language technologies; most are web-based anyway so they should work on your devices out of the box. After that, the pragmatics among you will brainstorm in small groups how the existing tools can be employed in your current practical needs and the visionaries among you will have a chance to propose new ideas for language processing tools for future implementation.

The workshop is **designed for all, who want to know more about what language technologies currently offer** and to get an insight into possibilities of their deployment **in the context of the work of the SAIs. Neither knowledge of computers technologies nor high-level proficiency in English is necessary.** Everything will be explained and the speech will be automatically subtitled.



4.2 Workshop worksheets

The participants of the workshop are given a set of workshop worksheets, available at <http://bit.ly/langtools-ws>. The worksheets contain links to the demoed tools, recapitulate the tasks that the participants should solve, and provide sample inputs for the tasks. As practically the same information is also included in the slides (see Section 4.3), we do not enclose the worksheet here.

4.3 Workshop slides

On the following pages, we enclose the slides prepared for the workshop, including back-up slides which are not shown to participants by default. The slides as shown to the participants can also be viewed online at <http://bit.ly/langtools-slides>.

LangTools Workshop

Benefit from natural language processing.

Ondřej Bojar, Vilem Zouhar, Tomáš Musil,
Michal Aueršperger, Rudolf Rosa, Dominik Macháček

Introduction and Workshop Overview

2

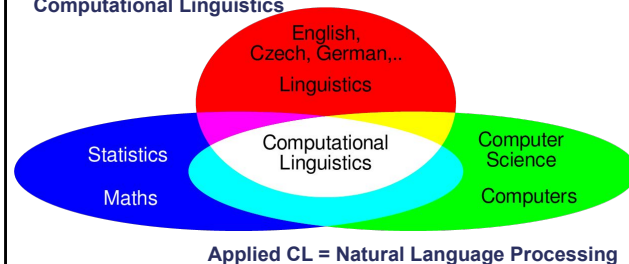
Where You Are and Why



Workshop on Automatic Processing
of Text and Speech
with potential use in your daily work, auditing practice

3

Computational Linguistics



4

Applied CL = Natural Language Processing

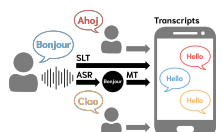
Text Input (T9), Spelling+Grammar Checking

gramar error are...

Internet Search, Information Extraction or Text Data Mining,
Sentiment Analysis, Text Summarization



Speech Recognition ("Speech to text"),
Machine Translation,
Spoken Language Translation



5

Text Input / Correction

T9 input method



gramar error are comon

6

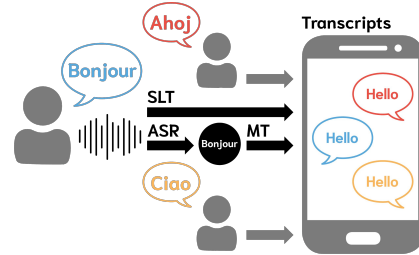
Internet Search, Text Analytics

Turning Big Data
into Useful Information



7

Multilingual Text and Speech Processing



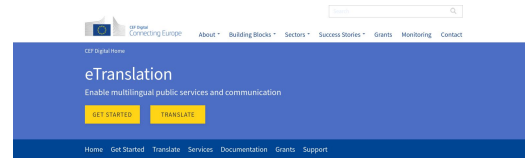
8

EU Support of NLP: Research -> Commodity: CEF, ELG

- The EU is well aware of the utility of NLP for the society.
- Many projects funded over the last decades.
- Commoditizing NLP
 - › European Language Grid
 - › eTranslation
 - › Other services

9

EU Support of NLP: Research -> Commodity

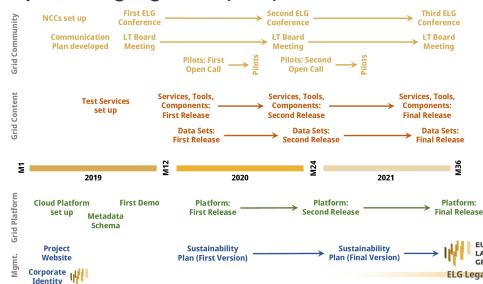


Latest News and Success Stories



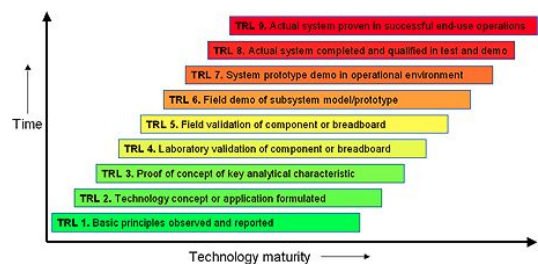
10

European Language Grid (ELG)



11

Technology Readiness Levels (TRL)



12

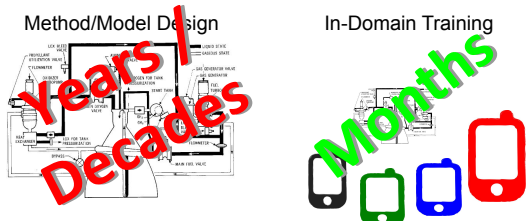
Training Data: TRL vs. Individual Domain and Language Support

- Most of the technologies are **trained**, i.e. some basic structure is **automatically populated**, relying on **language- and domain-specific manually annotated data**.
- Examples: translated audit reports, manually labelled named entities in German tax audit domain...
- So even a technology that achieves TRL 9 for Czech housing lease agreements may be not available or badly underperforming on Serbian road construction regulations.
- **Adapting** an existing technology to a new domain or language is **considerable work**, but it will take **months**, not **years** or **decades**.
- Here, we will showcase technologies that **are usually below TRL 9**, even in languages where they work best. ...we are a university, not a company.

13

TRL vs. Individual Domain and Language Support

- NLP components **are mostly trained**, not programmed.



14

Workshop Programme Blocks

- 1 › Presentation: [Linguistics and Language Processing](#)
› Demo: [Named Entity Recognition](#)
› Demo: [Question Answering](#)
- 2 › Demo: [Machine Translation](#)
› Demo: [Cross-Lingual Information Retrieval](#)
› Demo: [Speech Recognition & Speech Translation](#)
- 3 › Discussion: [Use language tools at your SAI!](#)

15

Your Ideas Are What Makes the Difference



16

Named Entity Recognition

17

DATES PEOPLE INSTITUTIONS LOCATIONS

Example 1

Input:

In January 2020 Daniel Hildegard together with members of the International Congress decided to visit Prague.

Output:

In **January 2020** **Daniel Hildegard** together with members of the **International Congress** decided to visit **Prague**.

18

Example 2

II. Subject of the Supplement No. 1

Since both **Contracting parties** are interested in continuing the relationship established by the Sublease agreement, they have agreed to extend the lease for a further **two years**, i.e. the lessee is entitled to use the apartment until **31st December 2020**. The other provisions of the Sublease agreement remain unchanged.

III. Final Provisions

In the event that any provision of this Supplement No. 1 is or it becomes invalid or ineffective, this shall not affect the validity or effectiveness of the other provisions of this Supplement No. 1.

In Art. III of the Sublease agreement, the tenant and the lessee agreed that the apartment in question would be rented to the tenant for a **fixed period** from **13th May 2016** to **31st December 2018**.

The Supplement No. 1 is bilingual. In the event of a dispute, the Czech version is decisive. The **Contracting parties** to this Supplement No. 1 declare that they have read the Supplement No. 1, agree with its contents and that the Supplement No. 1 was concluded freely, seriously, not in distress, under considerably unfavorable conditions.

In proof of these facts, both **parties** to this Agreement shall attach their handwritten signature.

Prague, **31st December 2018**

Karolína Černá, Lessee; **Marta Burešová**, Tenant

19

Task 1a - Who coordinated the audit?

Subject of audit: (CR) Excise duty administration (SR) Customs authorities procedures in excise duty administration.

The subject of the Agreement was cooperation in parallelly performed audits of „Excise duty administration“ included in the Audit plan of the SAO, CR for 2005, No. 05/34 and „Customs procedures in excise duty administration“ included in the SAO SR Audit plan for 2006, No. 48/11. Parallel audits had the nature of a coordinated audit. The cooperation consisted both, in the exchange of information that could not be obtained by Parties to the Agreement in course of excise duty administration audit on the territory of the respective state, and in drafting the joint final report on the result of audits in accordance with the European Implementing Guidelines for the INTOSAI Auditing Standard No 31.

The audit on the territory of the Czech Republic (hereinafter only „the CR“) was performed by audit teams composed of representatives of the State Budget Department and of the regional offices in the Central Bohemia, South Bohemia, West Bohemia, Northwest Bohemia, Northeast Bohemia, South Moravia, Central Moravia, and North Moravia from 7 February to 2 June 2006. The audit was performed by 27 auditors. One of the audited entities was the General Directorate of Customs (hereinafter only „the DGC“) and 11 tax offices. 12 out of 54 customs offices of the DGC were selected for the audit. The coordination was led by Reagan Johnston. The audit on the territory of the Slovak Republic (hereinafter only „the SR“) was performed by the Financial and Tax Section of the SAO SR in cooperation with the SAO sub-offices in Banská Bystrica and Košice in 5 out of the total number of 9 regional customs offices between 13 February and 2 June 2006.

The objective of parallel audits was to check procedures of the customs authorities in excise duty administration, as well as adherence to valid legislation in both countries and in the EU. The international cooperation was aimed at the procedures of tax authorities in supervising movement of excisable good.

20

Task 1b - What **institution** did the audit?

Each of both **Member States** is developing its own risk management system (hereinafter „RMS“), that has its strengths and weaknesses. Successful criteria, components or approaches of a RMS have to be exchanged and implemented in each **Member State** of the EC, if not so fraudsters will choose that **Member State** with the weakest RMS (see **5.**). As part of cooperation, the two **SAI** reviewed selected cases of intra-Community transactions processed by tax entities in the **CR** and **Germany**.

Thirty-one cases of business transactions were reviewed jointly using the legal provisions of the CLO, where there were doubts about their realization, their proper treatment or suspicion of VAT fraud. The **SAI** found that:

In some cases, the tax administration of another Member State refused to reply to a request for information. Some cases were detected, where taxpayers wrongly declared business transactions in their recapitulative statements. As a result, data in VIES were erroneous and therefore the tax administrators had to review those cases (see **6.**).

The audit was performed in the period from **June 2006** to **July 2006** by the **Division II – Department of State Budget, incomes** and by the territorial departments of **Central Bohemia**, **North-Western Bohemia**, **Southern Bohemia**, **Southern Moravia**, **Central Moravia** and **Northern Moravia**.

The audited entities were: the Ministry of Finance (hereinafter the „MoF“) and **10** tax offices – the tax office in **Humpolec**, the tax office in **Ústí nad Labem**, the tax office in **Kladno**, the tax office in **Liberec**, the tax office in **Nymburk**, the tax office in **Prostějov**, the tax office for **Prague 1**, the tax office for **Prague 4**, the tax office in **Sokolov** and the tax office in **Tábor**.

The conclusions of this audit was approved on **April 23, 2007**.

21

Named Entity Recognition

- information extraction of name entities
(= objects with proper names, such as Sony, John, Czechia)
- TRL 6/9
- search quickly through a document
- generate tags/topics for a document
- summarization (in keywords)
- automatic assignment of work (based on keywords)

22

Task 2 - practical

1. Go to: <https://explosion.ai/demos/displacy-ent>
2. Pick your preferred language (English, German, Spanish, Portuguese, ...)
3. Input a sentence with a named entity.
4. Click the search button. Examine the results.
5. Test this on larger texts, for example from Wikipedia.

23

Task 3 - practical

1. Go to: <https://explosion.ai/demos/displacy-ent>
2. Pick your preferred language (English, German, Spanish, Portuguese, ...)
3. Try to find an entity, which does not get classified.
Example: *WHO issued a new statement.*
However: **World Health Organization** issued a new statement.
However: **SAO** issued a new statement.
However: **NKU** issued a new statement.
4. Try to find an entity, which gets misclassified. What caused it?
Example: *I met with Kuba.*
However: *I met with John.*

24

Stanford Named Entity Tagger

Please enter your text here:

Only one existing cemetery in the vicinity, Fraser Cemetery in New Westminster (established in 1878), is older than Mountain View.

Submit Clear

Only one existing cemetery in the vicinity, **Fraser Cemetery** in **New Westminster** (established in 1878), is older than **Mountain View**.

Potential tags:

- ORGANIZATION
- LOCATION
- PERSON
- MISC

25

NamedTag

Model: czech-cnec2.0-140304

Input: ☒ Plain text ☐ Vertical

Output: ☒ XML (original text with annotations) ☐ Vertical (retrieved named entities only)

NKÚ má dnes v únoru workshop.

Process Input

Raw Output

Highlighted Output

NKÚ má dnes v **únoru** workshop.

26

displaCy Named Entity Visualizer

Národní kontrolní úřad is having a workshop with Institute of Formal And Applied Linguistics in Prague this February.

Model: English - en_core_web_sm (v2.2.0)

Entity labels (select all)

- ☒ PERSON ☒ NORR ☒ ORG
- ☒ GPE ☒ LOC ☒ PRODUCT
- ☒ EVENT ☒ WORK OF ART
- ☒ LANGUAGE ☒ DATE ☒ TIME
- ☒ PERCENT ☒ MONEY ☒ QUANTITY
- ☒ ORDINAL ☒ CARDINAL

Národní kontrolní úřad **ORG** is having a workshop with **Institute of Formal And Applied Linguistics** **ORG** in **Prague** **LOC** **this February** **DATE**.

27

Question Answering

28

Question Answering

- Imagine that you have a long document and you have a question.
- Question Answering software can find the answer for you!
- TRL: 6/7
- Task 1:**
 - > <https://zilinec.me/bert/>
 - > try the default context
 - ask: When will the conference happen?
 - ask: Who are the speakers?
 - ask: How much does it cost?

29

Question Answering – Task 2 – <https://zilinec.me/bert/>

Copy the following text into the context text area:

The audit was performed in the period from June 2006 to March 2007 by the Division II – Department of State Budget Incomes and by the territorial departments of Central Bohemia, North-Western Bohemia, Southern Bohemia, Southern Moravia, Central Moravia and Northern Moravia.

The audited entities were: the Ministry of Finance (hereinafter the "MoF") and 10 tax offices – the tax office in Humpolec, the tax office in Jihlava, the tax office in Kadan, the tax office in Liberec, the tax office in Nymburk, the tax office in Otrokovice, the tax office for Prague 1, the tax office for Prague 4, the tax office in Sokolov and the tax office in Třinec.

The conclusions of this audit was approved on April 23, 2007.

Ask questions, e.g.

- What institution did the audit?
- Who was audited?
- When was it finished?

30

Question Answering – Task 3 – <https://zilinec.me/bert/>

- random Wikipedia article:
<https://en.wikipedia.org/wiki/Special:Random>
- copy and paste the first paragraph into the tool
- ask questions

31

15 minutes break

Machine Translation

Main players:

- Google: www.translate.google.com
- Microsoft: www.bing.com/translator

Advantages:

- Available
- Free(*)
- Good results

34

Main players:

- Google: www.translate.google.com
- Microsoft: www.bing.com/translator

Disadvantages:


- Data protection
- General domain
- Uncommon language pair

35


Alternative services:


- LINDAT Translator:
<https://lindat.mff.cuni.cz/services/translation/>
- eTranslation:
<https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation>
- Ptakopet: <https://ptakopet.vilda.net/>
- ...


36




Altenmarkt-Zauchensee

Sponzorováno · 


Heute strahlend blauer Himmel in Altenmarkt.  Perfekte Bedingungen zum Langlaufen.

Dnes jasné modrá obloha v altenmarkt.  perfektní podmínky pro kříž lyžování.





Skryt originál · Ohodnotte tento překlad

37





Altenmarkt-Zauchensee

Sponzorováno · 

Heute strahlend blauer Himmel in Altenmarkt.  Perfekte Bedi...

zum Langlaufen.

Dnes jasné modrá obloha v altenmarkt.  perfektní podmínky pro kříž lyžování.



Skryt originál · Ohodnotte tento překlad

cross-country skiing

38

Google Translate

Text

Documents

CZECH · DETECTED

ENGLISH

SPANISH

FRENCH

GERMAN

SPANISH

ENGLISH


zitra jedu do osla


×


I'm going to the donkey tomorrow

☆

19/5000







Send feedback

39

Task #1:

Translating a part of an Estonian audit report:

See the [worksheet](#)

40

Task #2:

Using [Ptakopět](#) (3rd link in the worksheet), try to ask a clarifying question in Estonian

41

Ptakopět

bergamot

Source language: Czech

Target language: German

zitra jedu do osla

Ich fahre morgen nach Esel

Backward translation:

Paraphrases:

Zittra jedu do Osliku.

Enter experiment

42

CLIR: Cross-Lingual Information Retrieval

The problem

- Find **information**...
 - › in audits by **your SAI**
 - easy
 - › in audits by **other SAIs**
 - **sometimes** easy
 - EUROSAT database with English translations
 - usually **hard**

44

The solution

- **CLIR**: Cross-Lingual Information Retrieval
 - › search in **your language**
 - › find audits in **any language**
 - › read them in **your language**
- automatic translation
- TRL 5-7: demos and prototypes

45

CLIR demo

- **Languages**
 - › English (EN)
 - › Czech (CS)
 - › German (DE)
 - › French (FR)
- **Audits**
 - › Czech SAO (in Czech)
 - › Belgian SAI (in French)

46

CLIR Task 1: pension funds

- Go to bit.ly/ws-clir (see the worksheet), and find
 - › EN: documents relevant to **pension funds**
 - › CS: dokumenty týkající se **penzijních fondů**
 - › DE: Dokumente über die **Pensionsfonds**
 - › FR: des documents pertinent pour les **fonds de pension**

47



CLIR demo

Eurosal 2020 LangTools Workshop

English (EN)

Deutsch (DE)

Français (FR)

Česky (CS)

© 2020 Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czechia

22

ÚFAŁ

CLIR query

Search query:

Search

© 2020 Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czechia

22

ÚFAŁ

CLIR query

Search query:

Search

© 2020 Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czechia

22

ÚFAŁ

Results for query *pension funds*

Number of results found: 392

State pension funds

"...the **funds** concentrated in the special reserve account for the **pension** reform. The audit action was..."

Original name: Prostředky státu v oblasti důchodového pojištění

ID: K14008

Czech SAI (Nejvyšší kontrolní úřad), 2014

Czech, 16 pages, 6,854 words

Government financial assets, in particular those concentrated in the nuclear account

"...assets, in particular **funds** centered on a nuclear account The audit action was included in the SAO's..."

Original name: Státní finanční aktiva, zejména prostředky soustředěné na jaderném účtu

ID: K11009

Czech SAI (Nejvyšší kontrolní úřad), 2011

Czech, 13 pages, 5,636 words

report 'Comments and remarks on the draft state budget for the budgetary year 2017'

"...89 2 Programmes 25.54.6, 25.54.7 and 25.54.8 – Budget **funds** managed by Afsca on behalf of SPF..."

ID: 2016_49_Budget2017

Belgian SAI (Cour des comptes), 2016

French, 109 pages, 47,030 words

Government financial assets, in particular those concentrated in the nuclear account

"...assets, in particular **funds** centered on a nuclear account audit action was included in the SAO's..."

Original name: Státní finanční aktiva, zejména prostředky soustředěné na jaderném účtu

report 'Comments and remarks on the draft state budget for the budgetary year 2017'

"...89 2 Programmes 25.54.6, 25.54.7 and 25.54.8 – Budget funds managed by Afsca on behalf of SPF..."

Original name: rapport "Commentaires et remarques sur le projet de budget de l'Etat pour l'exercice budgétaire 2017"

ular those

ID: K11009

Czech SAI (Nejvyšší kontrolní úřad), 2011

Czech, 13 pages, 5,636 words

a nuclear account The

"

středky soustředěné na jaderném účtu

e draft state budget for

ID: 2016_49_Budget2017

Belgian SAI (Cour des comptes), 2016

French, 109 pages, 47,030 words

d 25.54.8 – Budget **funds**

es sur le projet de budget de l'Etat pour l'exercice budgétaire 2017"

ÚFAŁ

State pension funds

Highlighted for query: pension funds

Automatic translation

Original text

SAO Bulletin, control conclusions

Věstník NKÚ, kontrolní závěry

21

21

14/08

14/08

State **funds** in the field of pension insurance

Prostředky státu v oblasti důchodového pojištění

control action has been included in the control plan of the Supreme Audit Office (hereinafter referred to as 'SAO') for 2014 under „NKÚ“

Kontrolní akce byla zařazena do plánu kontrolní činnosti Nejvyššího kontrolního úřadu (dále jen „NKÚ“) na rok 2014 pod číslem 14/08. Kontrolní akce byla součástí kontrolního plánu NKÚ na rok 2014.

operation was managed and fidilia a kontrolní závěr

Kontrolní akce byla řízena a provedena v souladu s kontrolním plánem NKÚ na rok 2014.


14/08

Prostředky státu v oblasti důchodového pojištění

Kontrolní akce byla zařazena do plánu kontrolní činnosti Nejvyššího kontrolního úřadu (dále jen „NKÚ“) na rok 2014 pod číslem 14/08. Kontrolní akce byla součástí kontrolního plánu NKÚ na rok 2014.

© 2020 Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czechia

conclusions	zavery
21	21
14/08	14/08
State funds in the field of pension insurance control action has been included in the control plan of the Supreme Audit Office (hereinafter referred to as 'SAO') for 2014 under number 14/08. The audit operation was managed and	Prostředky státu v oblasti důchodového pojištění Kontrolní akce byla zařazena do plánu kontrolní činnosti Nejvyššího kontrolního úřadu (dále jen „NKÚ“) na rok 2014 pod číslem 14/08. Kontrolní akci řídila a kontrolní závěr



report 'Comments and remarks on the draft state budget for the budgetary year 2017'

Highlighted for query: pension funds

management schemes)353 is gestion globale)353 sont estimated at EUR 95 830,1 estimées à 95.830,1 millions million. The increase in d'euros. L'augmentation des expenditure of € 16.402,2 dépenses à raison de million (20.65%) is mainly 16.402,2 millions d'euros due to the integration of (20.65 %) s'explique surtout public pensions into social par l'intégration des pensions security expenditure. publiques dans les dépenses Effective April 1, 2016, the de la sécurité sociale. Depuis Federal Pension Service le 1er avril 2016, c'est le (FPS) pays these Service fédéral des pensions pensionss354. To this end, it (SFP) qui paye ces receives appropriations from pensions354. Il reçoit à cet the federal budget (entered in effet des dotations à la charge du Social Security SPF). du budget fédéral (inscrites au SPF Sécurité sociale355).

CHAPITRE III

Dépenses de la sécurité sociale


1 Évolution générale des dépenses

Dans le budget initial 2017, les dépenses consolidées de la sécurité sociale (CNSS-Gestion globale, Inasti-Gestion globale, Inasti-Soins de santé et les régimes hors gestion globale)⁽¹⁾ sont estimées à 95.830,1 millions d'euros. L'augmentation des dépenses à raison de 16.402,2 millions d'euros (20,65 %) s'explique surtout par l'intégration des pensions publiques dans les dépenses de la sécurité sociale. Depuis le 1^{er} avril 2016, c'est le Service fédéral des pensions (SFP) qui paye ces pensions⁽²⁾. Il reçoit à cet effet des dotations à la charge du budget fédéral (inscrites au SPF Sécurité sociale⁽³⁾).

Tableau 1 - Evolution des dépenses de la sécurité sociale (en millions d'euros)

© 2020 Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czechia

million (20.65%) is mainly due to the integration of public pension security expenditure. Effective April 1, 2016, the Federal Pension Service (FPS) pays these pensionss354. To this end, it receives appropriations from the federal budget (entered in the Social Security SPF).	16.402,2 millions d'euros (20,65 %) s'explique surtout par l'intégration des pensions publiques dans les dépenses de la sécurité sociale. Depuis le 1er avril 2016, c'est le Service fédéral des pensions (SFP) qui paye ces pensions354. Il reçoit à cet effet des dotations à la charge du budget fédéral (inscrites au SPF Sécurité sociale355).
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



Resultate für die Suchanfrage Pensionsfonds

Anzahl der gefundenen Resultaten: 13

Bericht 'Die Pensionsmaschine: Entwicklung und Anwendung von Versorgungsleistungen im öffentlichen Dienst'

ID: 2018_41_MoteurPension
belgische SAI (Cour des comptes), 2018
französisch, 30 Seiten, 12.866 Wörter

"...Datum", d.h. das frühestmögliche Rentendatum. Das Modul zur Berechnung des P-Datums des Pensionsfonds..."

Originale Name: rapport "Le moteur des pensions: élaboration et application pour les pensions de la fonction publique"

report 'Implementation of the Capelo project and processing of electronic data by the Federal Pensions Service - civil service pensions' (document written in French)

ID: 2017_07_MiseOeuvreCapelo
belgische SAI (Cour des comptes), 2017
französisch

"...teilweise vom Pensionsfonds des öffentlichen Dienstes auf die Personalabteilung jedes öffentlichen..."

Bericht 'Renten mit ausländischer Komponente'

ID: 2016_13_Pensions



Staatliche Vermögenswerte, insbesondere solche, die in der Nuklearrechnung konzentriert sind

transfèrent. uctu statni pokladny.

Zinserträge aus Anlagen Ürokový příjem z sowohl von Kernfonds als investování prostředků auch von Pensionsfonds sind jederného i důchodového Einnahmen aus dem účtu je příjmem státního Staatshaushalt des OSFA-rozpčtu kapitoly OSFA a Kapitels und sind současně je výdajem státního gleichzeitig Ausgaben des rozpočtu v kapitole Státní Staatshaushalts im Kapitel dluh. Při investování repo Staatsschulden. Bei der operacemi se peněžní Investition in Repos werden prostředky převádějí na účty Gelder auf kommerzielle Bankkonten überwiesen. Die investované prostředky je so investierten Mittel senken den Fondsbestand im prostředků na souhrnném allgemeinen Konto des účtu státní pokladny, tím se

Investování peněžních prostředků do státních dluhopisů má charakter „kvasitizování“: dochází k převodu těchto prostředků mezi účty podřízenými souhrnnému účtu státní pokladny. Úrokový příjem z investování peněžních prostředků je příjmem státního rozpočtu kapitoly OSFA a současně je výdajem státního rozpočtu v kapitole Státní dluh. Při investování repo operacemi se peněžní prostředky převádějí na účty komerčních bank. O tomto investovaném prostředku je v účtu státní pokladny na souhrnném účtu státní pokladny, tím se zvyšují náklady na zaplacení dle státní pokladny včetně úrokových výnosů.

Kapitole OSFA se na jiných rozpočtových kapitolách liší tím, že neobsahuje rozpočtové příjmy a výdaje správy kapitálu, ale podle zákona o rozpočtových pravidlech j) tvoří peněžní operace na účtech SFA a výjimkou operací spojených s obstaráváním a umisťováním státního dluhu. V roce 2018 představovaly skutečné výnosy kapitoly OSFA částku 1 529,1 mil. Kč, tj. 43,65 % schváleného rozpočtu, v roce 2019 to bylo 201,7 mil. Kč, tj. pouze 10,90 % schváleného rozpočtu. U posledních dat výpočtových položek kapitoly OSFA nedostatečně najím výnosům, ale k transferům do jiných kapitol státního rozpočtu prostřednictvím rozpočtových operací, které dlouhodobě tvoří v kapitole OSFA náležit finanční výnosy jiných rozpočtových kapitol.

V kontrolovaném období byly poskytovány zejména obcím ze SFA „mimořádné dotace“ (v roce 2009 ve výši 166,1 mil. Kč a v roce 2010 ve výši 87,48 mil. Kč), které fakticky tvořily „skrytou“ rozpočtovou rezervu pro MF.

O neoprávněném nastavení rozpočtového procesu v kapitole OSFA svědčí i skutečnost, že v roce 2009 (2010) nebyly peníze poskytnuty dle předpokládaných výnosů této kapitoly MF část příjmů a výnosů na účtech SFA v kapitole OSFA neoprávněně, část příjmů zahrnovala

CLIR Task 2: family reunification

- Go to bit.ly/ws-clir (see the worksheet), and find
 - EN: Number of Belgian visas for **family reunification** in 2018
 - CS: Počet belgických víz pro **sloučení rodiny** v roce 2018
 - DE: Anzahl der belgischen Visa für die **Familienzusammenführung** im Jahr 2018
 - FR: Nombre de visas belges pour le **regroupement familial** en 2018

UFA Results for query *family reunification*
Number of results found: 84

<p>press release 'Belgian Immigration Office : Processing Applications for Family Reunification'</p> <p>"...Report to the Federal Parliament: Office for Aliens: processing of applications for <i>family reunification</i>..."</p> <p>Original name: communiqué de presse "Office belge de l'immigration: traitement des demandes de regroupement familial"</p>	<p>ID: 2020_02_RegroupementFamilial_Communique Belgian SAI (Cour des comptes), 2020 French, 2 pages, 818 words</p>
<p>report 'Belgian Immigration Office : Processing Applications for Family Reunification'</p> <p>"...for <i>family reunification</i> Office for Aliens: processing of applications applications for <i>family</i>..."</p> <p>Original name: rapport "Office belge de l'immigration: traitement des demandes de regroupement familial"</p>	<p>ID: 2020_02_RegroupementFamilial Belgian SAI (Cour des comptes), 2020 French, 60 pages, 17,779 words</p>
<p>report 'Full Unemployment Benefits – Prevention and Detection of Undue Payments'</p> <p>"...Principles of supervision of <i>family</i> categories 24.3.2.2</p>	<p>ID: 2018_04_AllocationChomageComple Belgian SAI (Cour des comptes), 2018 French, 56 pages, 15,807 words</p>

UFA report 'Belgian Immigration Office : Processing Applications for Family Reunification'
Highlighted for query: family reunification

<p>Aliens Office: processing of applications <i>family reunification</i></p> <p>In 2018, 13,946 visas were issued for <i>family reunification</i>, representing 43% of long-term visas.</p> <p>Of these, 311 visas were issued without examination of the application due to an overstay.</p> <p>In the same year, the Aliens Office received 83,932</p>	<p>Office des étrangers : traitement des demandes de regroupement familial</p> <p>En 2018, 13,946 visas ont été délivrés en vue d'un regroupement familial, soit 43 % des visas de longue durée.</p> <p>Parmi ceux-ci, 311 visas ont été délivrés sans examen de la demande, en raison d'un dépassement de délai.</p> <p>Au cours de la même année, l'Office des étrangers a reçu 83 932 demandes de séjour</p>	<p>Office des étrangers : traitement des demandes de regroupement familial</p> <p>En 2018, 13 946 visas ont été délivrés en vue d'un regroupement familial, soit 43 % des visas de longue durée.</p> <p>Parmi ceux-ci, 311 visas ont été délivrés sans examen de la demande, en raison d'un dépassement de délai.</p> <p>Au cours de la même année, l'Office des étrangers a reçu 83 932 demandes de séjour (nouvelles demandes ou demandes de prolongation).</p> <p>Le regroupement familial est une procédure qui permet aux personnes étrangères dont un membre de la famille réside en Belgique de venir le rejoindre à certaines conditions. Ce</p>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

© 2020 Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czechia

Aliens Office: processing of applications <i>family reunification</i>	Office des étrangers : traitement des demandes de regroupement familial
In 2018, 13,946 visas were issued for <i>family reunification</i> , representing 43% of long-term visas.	En 2018, 13.946 visas ont été délivrés en vue d'un regroupement familial, soit 43 % des visas de longue durée.
Of these, 311 visas were issued without examination of the application due to an	Parmi ceux-ci, 311 visas ont été délivrés sans examen de la demande, en raison d'un

CLIR Task 3: look for some other information

• Go to bit.ly/ws-clir (see the worksheet), and look e.g. for:

- crime prevention
 - › prevence kriminality
 - › Kriminalprävention
 - › prévention du crime
- highways
 - › dálnice
 - › Autobahnen
 - › autoroutes
- state budget
 - › státní rozpočet
 - › Staatshaushalt
 - › budget de l'État
- ...

64

Automatic Speech Recognition & Spoken Language Translation

Automatic Speech Recognition & Speech Translation

Automatic Speech Recognition (ASR)
Machine Translation (MT)
Spoken Language Translation (SLT)

Transcript = what was said in the same language
Translation = what was said in other language

66

Cross-lingual communication

TRL: 5/6



67

10 minutes break

Discussion Groups

Discussion groups

- Form groups of 4-6 people
- How could language tools could **help** at **your SAI**?
 - › discuss **possible uses** of language tools (at least 3)
 - › **suggest new** useful language tools (at least 3)
 - › tools can be adapted, improved, invented...
- Our **facilitators** will help you!
 - › they can suggest some **ideas**
 - › they can **estimate** if your ideas could work
- At the end, each group **summarizes** their ideas

70

THANK YOU FOR YOUR
KIND ATTENTION

CZECH REPUBLIC
SUPREME AUDIT OFFICE
www.nku.cz



5 Conclusion

This deliverable presented the LangTools NLP workshop prepared for the EUROSAT Congress in June 2020.

The workshop mostly consists of interactive demo sessions, showing various NLP tools which could potentially be useful for SAT auditors. The participants are actively encouraged to try out the tools themselves and assisted by experienced facilitators.

All the materials have been prepared and the workshop has been rehearsed several times. Due to Covid-19, the congress and the workshop have not taken place yet and are currently planned for June 2021.