



LangTools for ERSTE

Benefit from natural language processing

ÚFAL

Ondřej Bojar,
Tomáš Musil, Rudolf Rosa,
Jindřich Libovický, Matúš Žilinec

Introduction and Workshop Overview

Workshop Outline

- Intro: Linguistics and Language Processing
 - Ondrej, 10 min
- Demo throughout the session: Speech Recognition and Translation
- Presentation/Demo: Immediate Response Machine Translation
 - Jindrich, 15 min
- Demo: Cross-Lingual Information Retrieval
 - Ruda, 20 min
- Demo: Question Answering
 - Tomáš, 20 min
- Presentation: Towards Automatic Summarization of Meetings
 - Ondrej, 10 min
- Discussion: Ideas for future collaboration
 - everyone, 20 min

Where You Are and Why

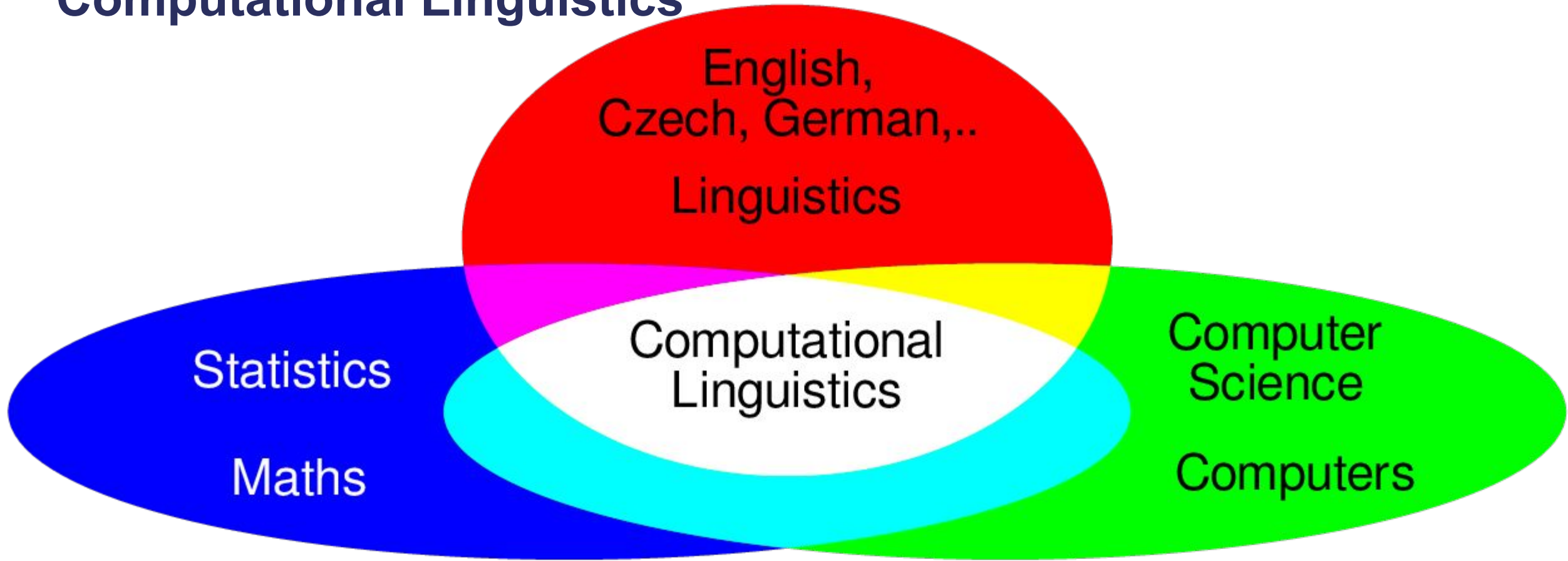


**Workshop on Automatic Processing
of Text and Speech**
*with potential for future collaboration
between ERSTE and ÚFAL*

Your Ideas Are What Makes the Difference



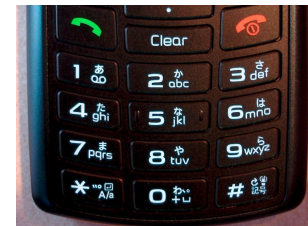
Computational Linguistics



Applied CL = Natural Language Processing

Applied CL = Natural Language Processing

Text Input (T9), Spelling+Grammar Checking

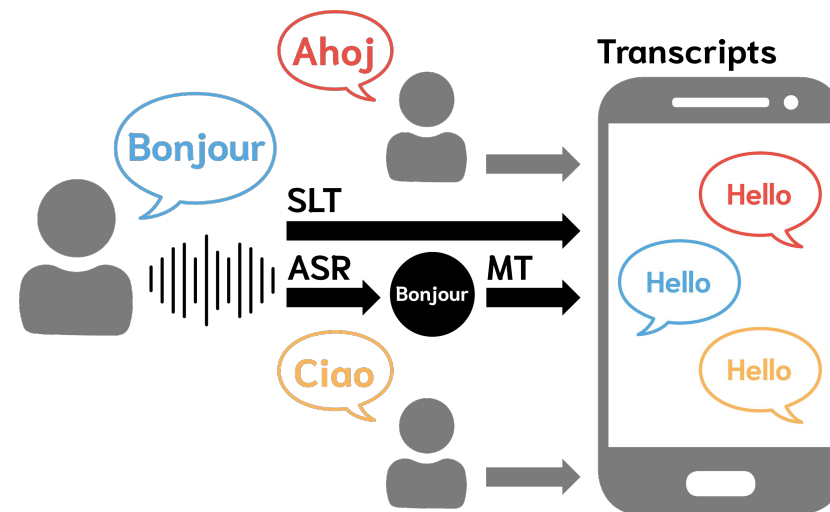


gramar error are...

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



Speech Recognition (“Speech to text”),
Machine Translation,
Spoken Language Translation



Text Input / Correction

T9 input method



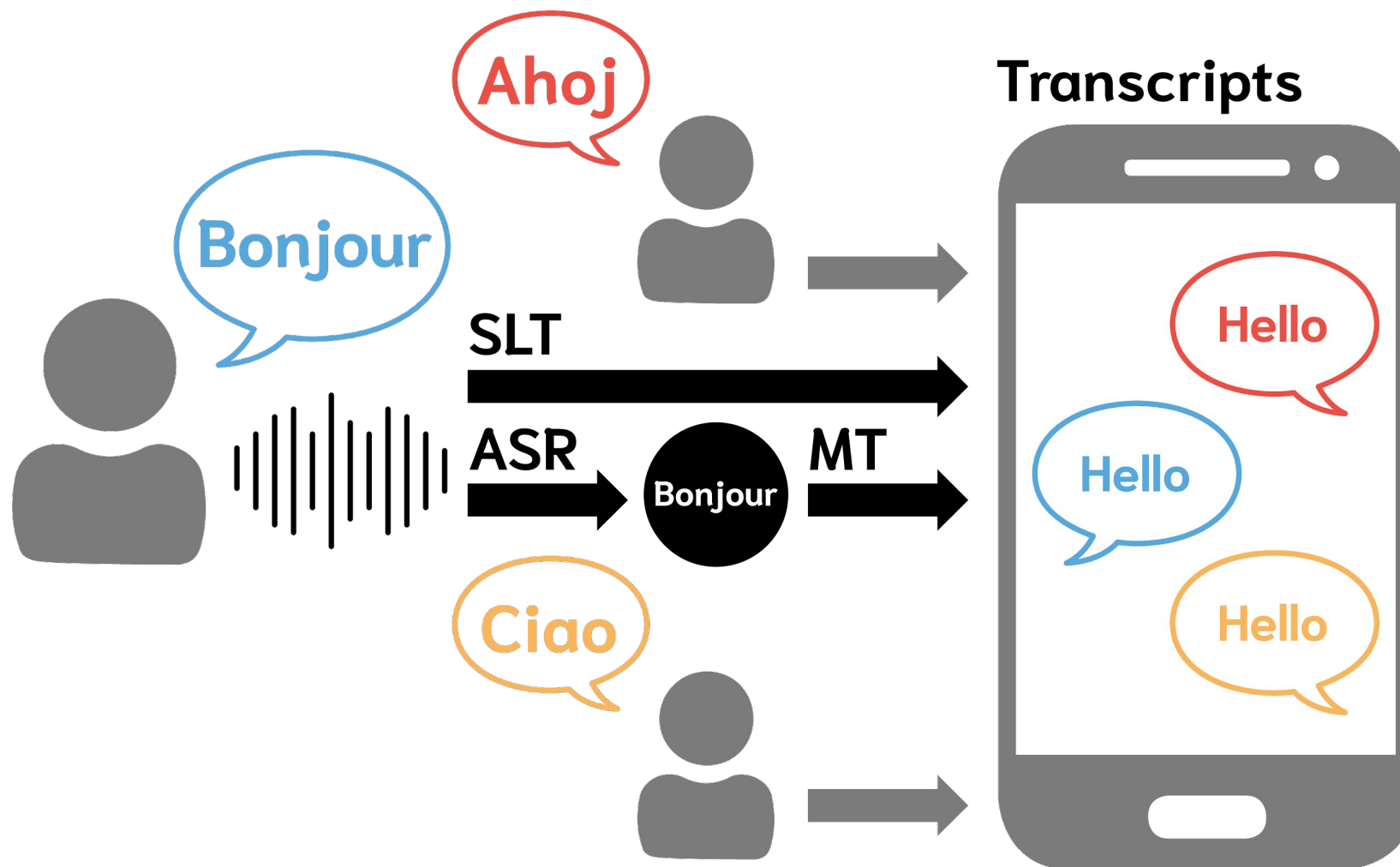
gramar error are comon

Internet Search, Text Analytics

Turning Big Data
into Useful Information



Multilingual Text and Speech Processing



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

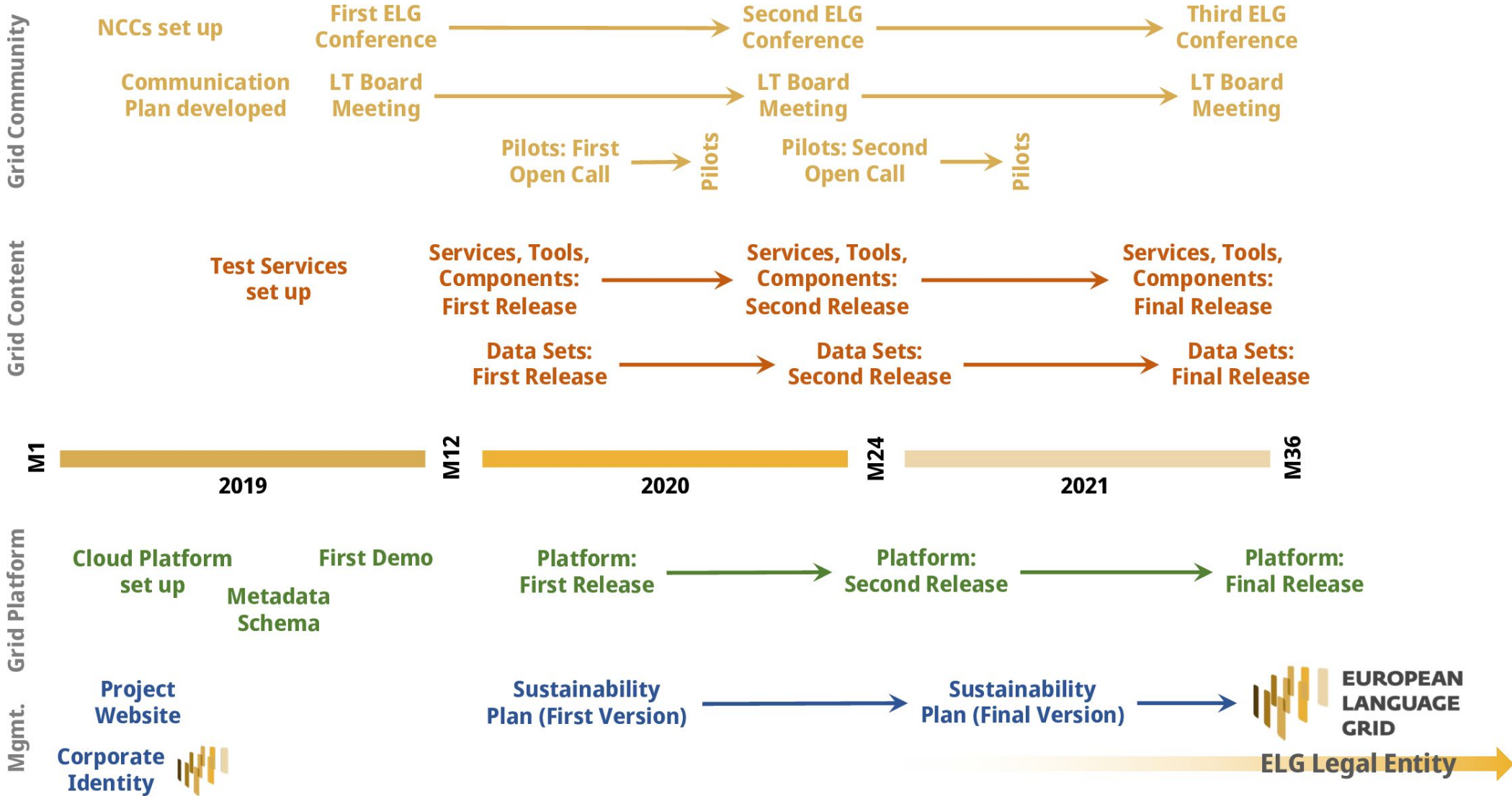
EU Support of NLP: Research -> Commodity

The screenshot shows the CEF Digital Home website. At the top right is a search bar with the text "Search" and a magnifying glass icon. Below it is a navigation menu with the following items: "About", "Building Blocks", "Sectors", "Success Stories", "Grants", "Monitoring", and "Contact". The main content area has a blue background and features the text "CEF Digital Home" and "eTranslation" in large white font. Below "eTranslation" is the subtitle "Enable multilingual public services and communication". There are two yellow buttons: "GET STARTED" and "TRANSLATE". At the bottom of the page is a dark blue footer with white text: "Home", "Get Started", "Translate", "Services", "Documentation", "Grants", and "Support".

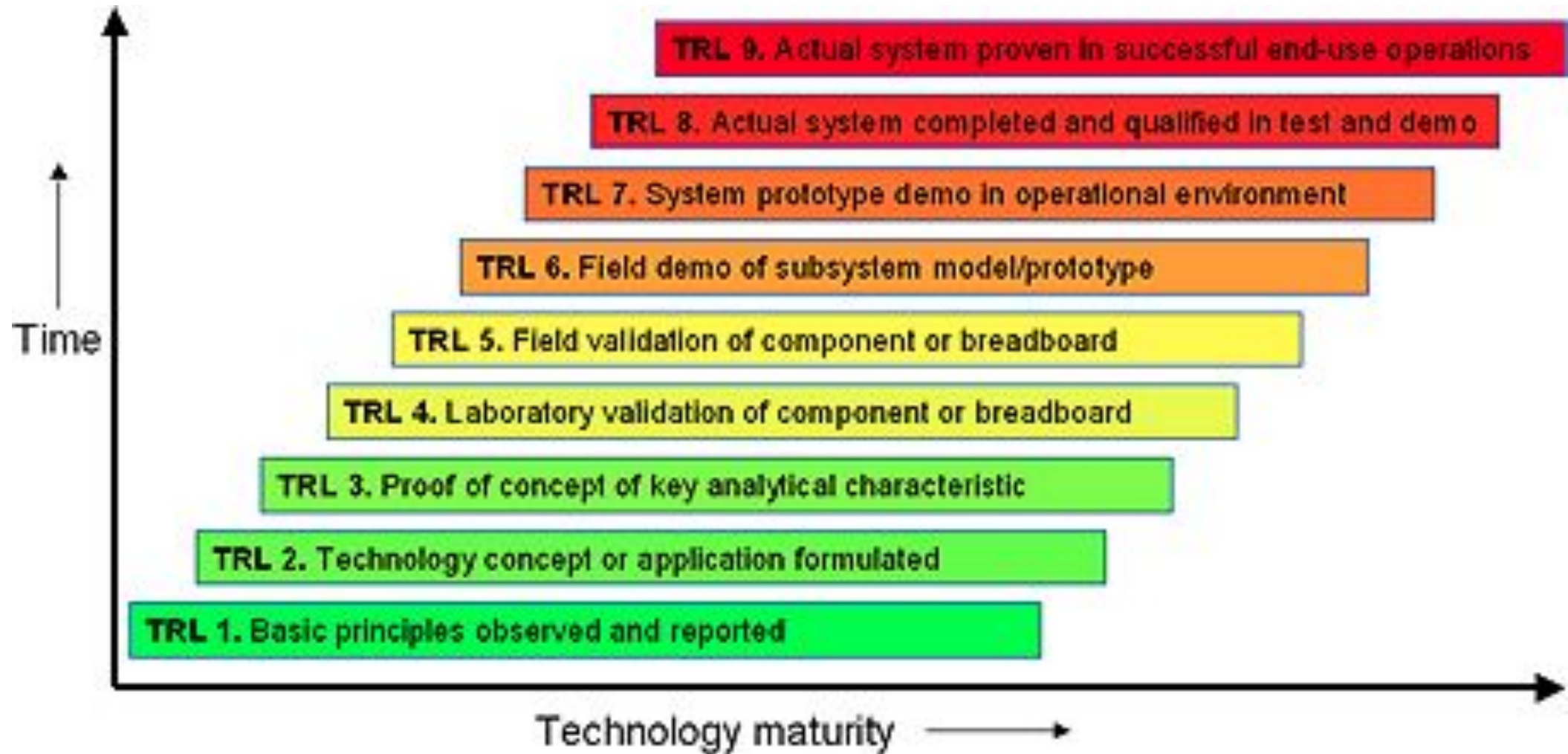
Latest News and Success Stories



European Language Grid (ELG)



Technology Readiness Levels (TRL)



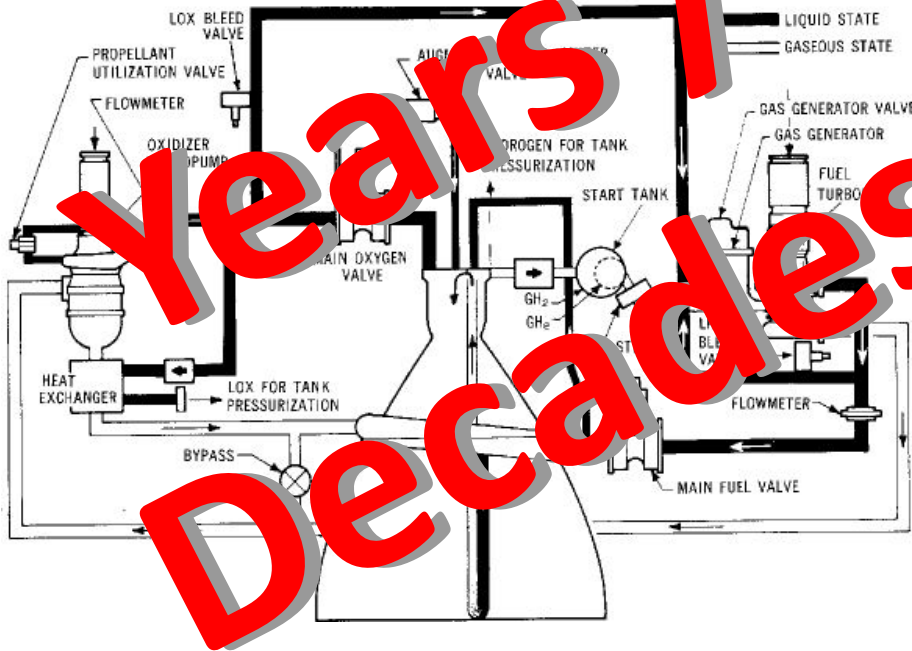
Training Data: TRL vs. Individual Domain and Language Support

- 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.

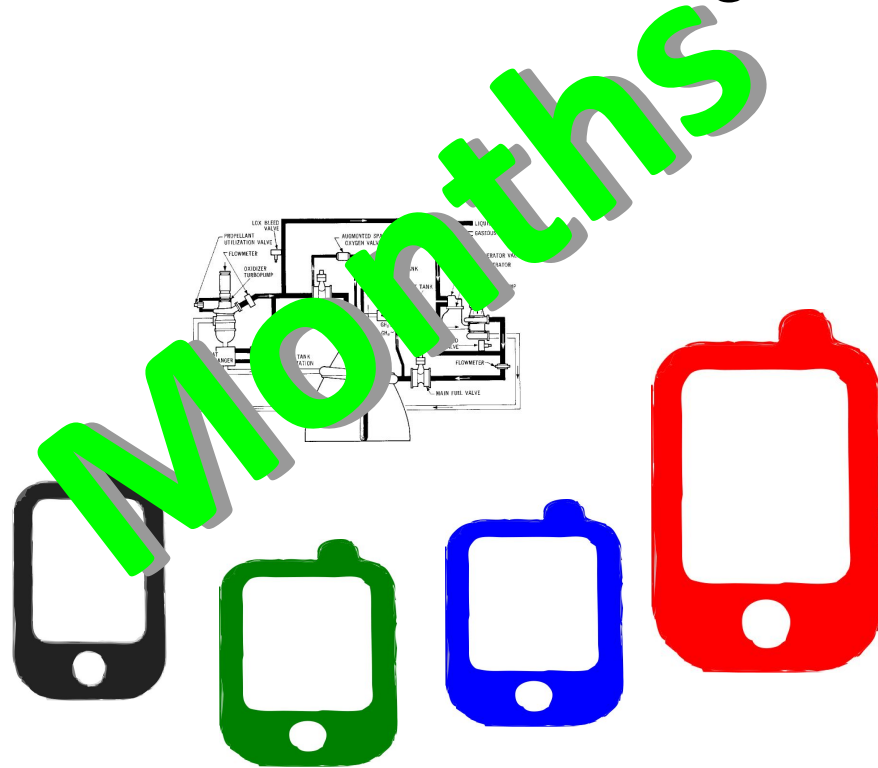
TRL vs. Individual Domain and Language Support

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

Method/Model Design



In-Domain Training

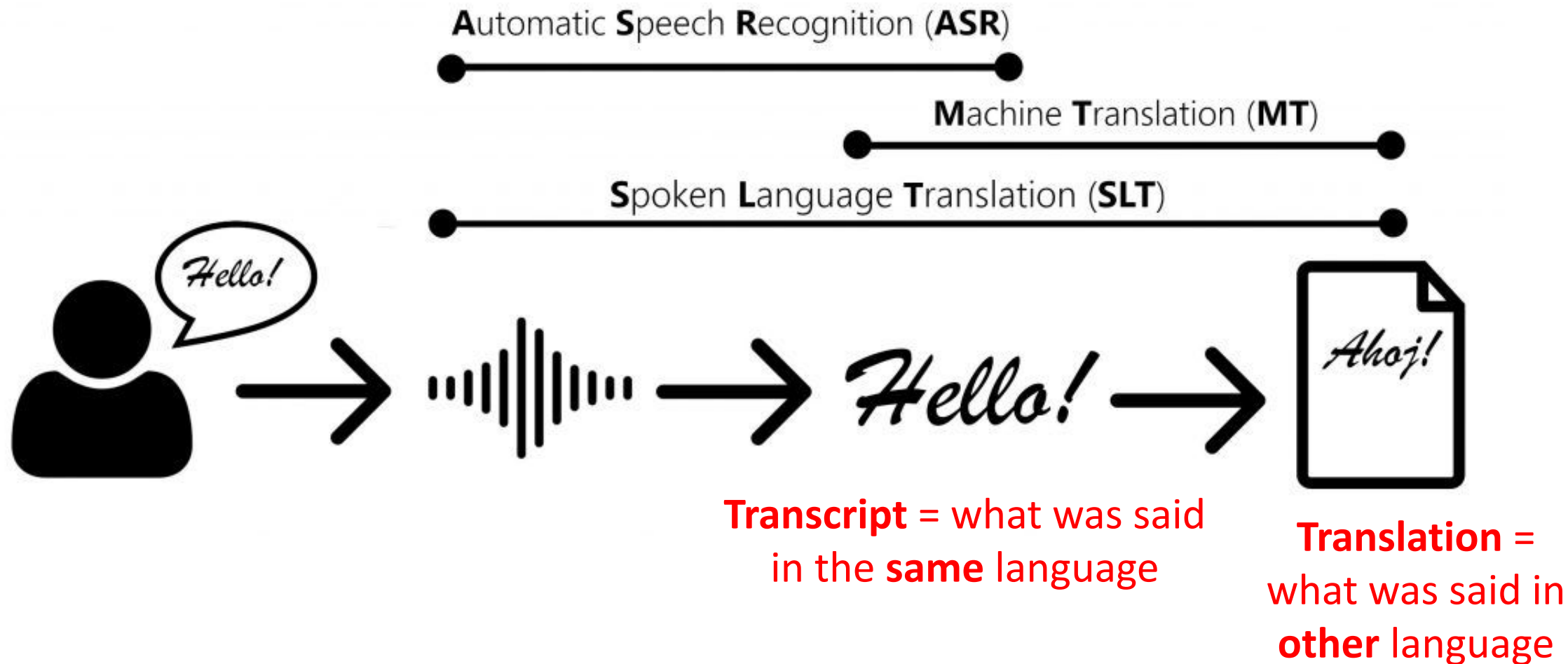


Automatic Speech Recognition & Speech Translation

URL to watch throughout the call

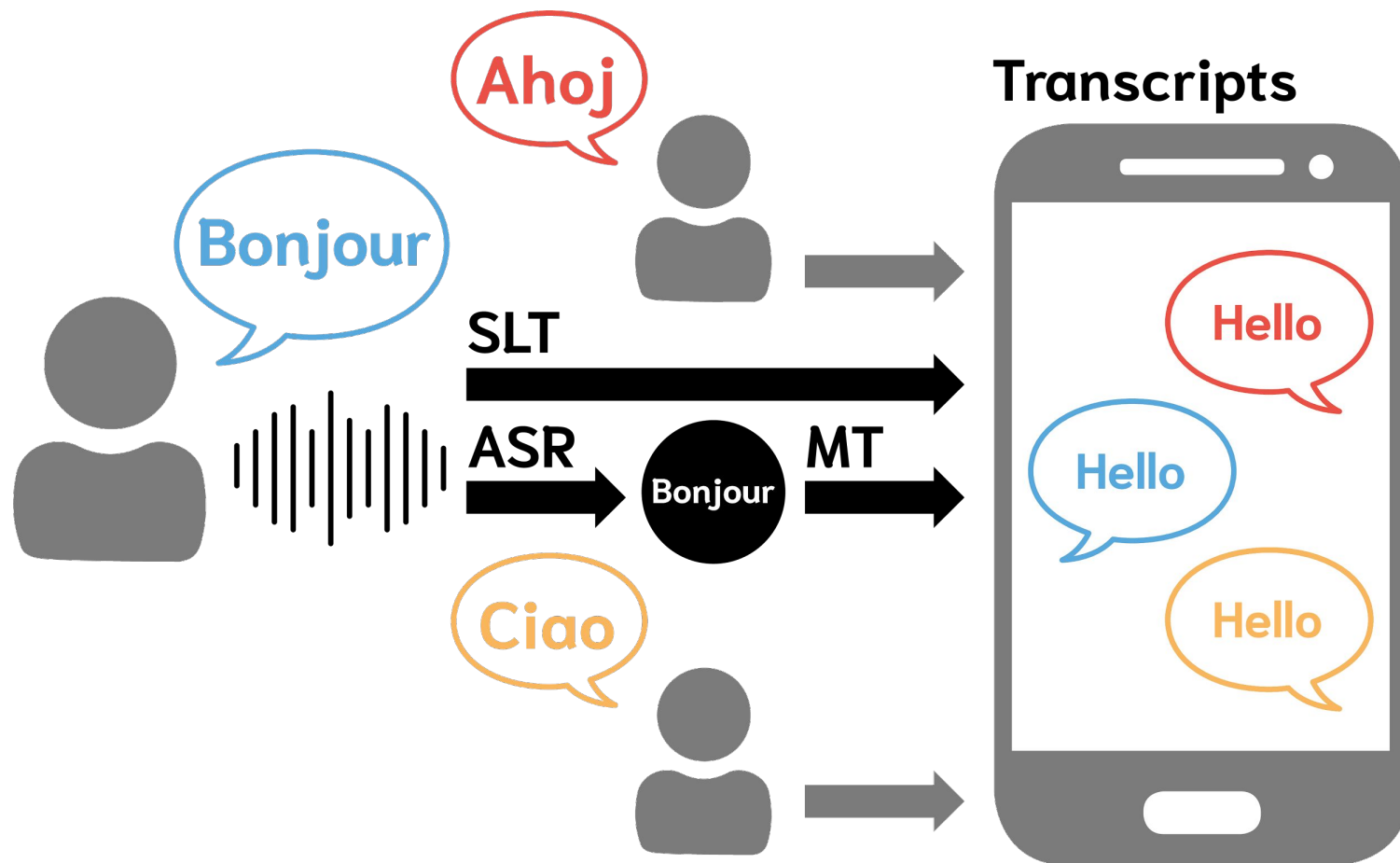
<https://quest.ms.mff.cuni.cz/elitr/demo/>

Automatic Speech Recognition & Speech Translation



Cross-lingual communication

TRL: 5/6



<https://elitr.eu/blog/>

📅 5. 11. 2020 👤 tereza 💬 Leave a Comment

Everything can go wrong in live subtitling

Humans are a very intricate species. Many activities were perceived to be human-specific for a long time. These would be for example playing board games, driving cars, or even artistic endeavours like painting or composing music. Many technological breakthroughs happened since the emergence of computers and deep neural networks and suddenly these activities cannot be perceived as limited to humans anymore.

Surely we had to pick the examples of “no-longer-human-specific” activities tendentiously! There are other activities such as those that distinguish our species from animals. Those are the true human activities that will possibly never be executed on the same level by computers — like human speech.

Computational linguistics

There are many research groups that try to automate language processing to human perfection. Thanks to these efforts, with a targeted focus and most modern methods exercised for over a decade, English-Czech text translation reached the quality of humans in some areas (<https://www.nature.com/articles/s41467-020-18073-9>). This success builds upon superfast computers and very large amounts of data. New data for computational linguistics is basically being created every second. People communicate all the time.



Objectives

ASR: [Automatic Speech Recognition](#)

SLT: [Spoken Language Translation](#)

MT: [Machine Translation](#)

AM: [Automatic Minuting](#)

Immediate Response Machine Translation

<https://docs.google.com/presentation/d/1hENadGT45fN8EBHXHG41tbjTbAOC8doNwBhNtq8U0Qc/edit?usp=sharing>

<https://ufal.mff.cuni.cz/ufal-ukraine>

Named Entity Recognition

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.

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

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.

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 Jihlava, the tax office in Kadaň, 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.

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)

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.

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: *NKÚ 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.*

Stanford Named Entity Tagger

Please enter your text here:

Only one existing cemetery in the vicinity, Fraser Cemetery in New Westminster (established in 1870), is older than Mountain View.

Only one existing cemetery in the vicinity, Fraser Cemetery in New Westminster (established in 1870), is older than Mountain View.

Potential tags:

ORGANIZATION

LOCATION

PERSON

MISC

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.

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)

<input checked="" type="checkbox"/> PERSON	<input checked="" type="checkbox"/> NORP	<input checked="" type="checkbox"/> ORG
<input checked="" type="checkbox"/> GPE	<input checked="" type="checkbox"/> LOC	<input checked="" type="checkbox"/> PRODUCT
<input checked="" type="checkbox"/> EVENT	<input checked="" type="checkbox"/> WORK OF ART	
<input checked="" type="checkbox"/> LANGUAGE	<input checked="" type="checkbox"/> DATE	<input checked="" type="checkbox"/> TIME
<input checked="" type="checkbox"/> PERCENT	<input checked="" type="checkbox"/> MONEY	<input checked="" type="checkbox"/> QUANTITY
<input checked="" type="checkbox"/> ORDINAL	<input checked="" type="checkbox"/> CARDINAL	

Národní Kontrolní Úřad **ORG** is having a workshop with Institute of Formal And Applied Linguistics **ORG** in Prague **GPE** this February **DATE** .

Machine Translation

Main players:

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

Advantages:

- Available
- Free(*)
- Good results

Main players:

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

Disadvantages:

- Data protection
- General domain
- Uncommon language pair

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/>


- ...



Altenmarkt-Zauchensee



Sponzorováno · 

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

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

 · [Skrýt originál](#) · [Ohodnoťte tento překlad](#)



Altenmarkt-Zauchensee

Sponzorováno ·

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

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

· [Skrýt originál](#) · [Ohodnoťte tento překlad](#)

cross-country skiing



🗨️ Text

📄 Documents

CZECH - DETECTED

ENGLISH

SPANISH

FRENCH



GERMAN

SPANISH

ENGLISH



zitra jedu do osla|



I'm going to the donkey tomorrow



18/5000



[Send feedback](#)

Task #1:

- Translating a part of an Estonian audit report:

See the [worksheet](#)

Task #2:

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



Source language: Czech

Target language: German

zitra jedu do osla

Ich fahre morgen nach Esel

Backward translation:

Paraphrases:

Zítřa jedu do Oslíku.

Enter experiment

CLIR: Cross-Lingual Information Retrieval

The problem

- Find **information...**
 - › in documents produced by **your organization branch**
 - should be easy
 - › in documents produced by **other organization branches**
 - sometimes **easy**
 - sometimes **hard**
 - language barrier

The solution

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

CLIR demo at bit.ly/ws-clir

- **Demo languages**

- › English (EN)
- › Czech (CS)
- › German (DE)
- › French (FR)

- **Demo data: audits from Supreme Audit Institutions**

- › Czech SAO (in Czech)
- › Belgian SAI (in French)

- Apache Solr information retrieval + Lindat machine translation

CLIR Task 1: pension funds

- **Go to bit.ly/ws-clir 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**



CLIR demo

Eurosai 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



CLIR query

Search query:

Search

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



CLIR query

Search query:

Search

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



Results for query *pension funds*

Number of results found: 392

State pension funds

ID: K14008

Czech SAI (Nejvyšší kontrolní úřad), 2014
Czech, 16 pages, 6,854 words

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

Original name: Prostředky státu v oblasti důchodového pojištění

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

ID: K11009

Czech SAI (Nejvyšší kontrolní úřad), 2011
Czech, 13 pages, 5,636 words

"...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

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

ID: 2016_49_Budget2017

Belgian SAI (Cour des comptes), 2016
French, 109 pages, 47,030 words

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

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 soustředě

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 managed by Afsca on behalf of SPF..."

Original name: rapport "Commentaires et remarques sur le projet de

cular those

ID: K11009

Czech SAI (Nejvyšší kontrolní úřad), 2011

a nuclear account The

Czech, 13 pages, 5,636 words

"

středky soustředěné na jaderném účtu

e draft state budget for

ID: 2016_49_Budget2017

Belgian SAI (Cour des comptes), 2016

d 25.54.8 – Budget *fonds*

French, 109 pages, 47,030 words


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



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 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

Věstník NKÚ, kontrolní závěry  21

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 akci řídila a kontrolní závěr vypracovala členka NKÚ JUDr. Eliška Kadaňová.

Cílem kontroly bylo prověřit vykazované údaje z oblasti výběru pojistného na důchodové pojištění a údaje z oblasti výplaty dávek důchodového pojištění; prověřit, zda Ministerstvo financí při správě prostředků soustředěných na zvláštním účtu rezervy pro důchodovou reformu postupovalo v souladu s právními předpisy.

Kontrolní akce byla schválena v návaznosti na usnesení vlády ČR¹, kterým dala vláda podnět NKÚ zaměřit se na prověření tzv. zvláštního účtu rezervy pro důchodovou reformu. Po provedeném rozboru byl podnět vlády ČR rozšířen.

Kontrola byla prováděna v době od března do září roku 2014. Kontrolovaným obdobím byly roky 2009–2013, v případě věcných souvislostí i období předcházející a následující.

Kontrolované osoby:

- Ministerstvo práce a sociálních věcí,

conclusions

zavery

21

21

14/08

14/08

State **funds** in the field of **pension** insurance. The control action has been included in the control plan of the Supreme Audit Office (hereinafter referred to as

Prostředky státu v oblasti důchodového pojištění. Kontrolní akce byla zařazena do plánu kontrolní činnosti Nejvyššího kontrolního úřadu (dále jen

'SAO') for 2014 under number 14/08. The audit operation was managed and

„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 **pension**s 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 **pension**s³⁵⁴. 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 the 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 (ONSS-Gestion globale, Inasti-Gestion globale, Inami-Soins de santé et les régimes hors gestion globale)³⁵³ 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 pensions³⁵⁴. Il reçoit à cet effet des dotations à la charge du budget fédéral (inscrites au SPF Sécurité sociale³⁵⁵).

Tableau 1 – Évolution des dépenses de la sécurité sociale (en millions d'euros)

million (20.65%) is mainly due to the integration of public pensions into social security expenditure. Effective April 1, 2016, the Federal Pension Service (FPS) pays these pensions³⁵⁴. 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 Resultate: 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

transferiert. uctu statni pokladny.
Zinserträge aus Anlagen Úrokový příjem z
sowohl von Kernfonds als investování prostředků
auch von Pensionsfonds sind jaderného i důchodového
Einnahmen aus dem účtu je příjem státního
Staatshaushalt des OSFA- rozpočtu 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 komerčních bank. O takto
Bankkonten überwiesen. Die investované prostředky je
so investierten Mittel senken nižší stav peněžních
den Fondsbestand im prostředků na souhrnném
allgemeinen Konto des účtu státní pokladny, tím se

Investování peněžních prostředků do státních dluhopisů má charakter „kvaziinvestování“; dochází k převodům těchto prostředků mezi účty podřízenými souhrnnému účtu státní pokladny. Úrokový příjem z investování prostředků jaderného i důchodového účtu 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 takto investované prostředky je nižší stav peněžních prostředků na souhrnném účtu státní pokladny, tím se zvyšují nároky na zajištění denní likvidity státní pokladny včetně úrokových výdajů.

Kapitola OSFA se od jiných rozpočtových kapitol liší tím, že neobsahuje rozpočtové příjmy a výdaje správce kapitoly, ale podle zákona o rozpočtových pravidlech ji tvoří peněžní operace na účtech SFA s výjimkou operací spojených s obsluhou a umořováním státního dluhu. V roce 2009 představovaly skutečné výdaje kapitoly OSFA částku 1 539,1 mil. Kč, tj. 43,65 % schváleného rozpočtu, v roce 2010 to bylo 261,7 mil. Kč, tj. pouze 10,90 % schváleného rozpočtu. U podstatné části výdajových položek kapitoly OSFA nedocházelo k reálným výdajům, ale k transferům do jiných kapitol státního rozpočtu prostřednictvím rozpočtových opatření, která dlouhodobě tvoří v kapitole OSFA nástroj financování výdajů 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 97,48 mil. Kč), které fakticky tvořily „skrytou“ rozpočtovou rezervu pro MF.

O nesprávném nastavení rozpočtového procesu v kapitole OSFA svědčí i skutečnost, že v roce 2009 i 2010 nebyla čerpána podstatná část předpokládaných výdajů této kapitoly.

MF část příjmů a výdajů na účtech SFA v kapitole OSFA nevykazovalo, část příjmů zahrnovalo

CLIR Task 2: family reunification

- **Go to** bit.ly/ws-clir 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



Results for query *family reunification*

Number of results found: 84

press release 'Belgian Immigration Office : Processing Applications for Family Reunification'

"...Report to the Federal Parliament: Office for Aliens: processing of applications for *family reunification*..."

Original name: communiqué de presse "Office belge de l'immigration: traitement des demandes de regroupement familial"

ID: 2020_02_RegroupementFamilial_Communique
Belgian SAI (Cour des comptes), 2020
French, 2 pages, 818 words

report 'Belgian Immigration Office : Processing Applications for Family Reunification'

"...for *family reunification* Office for Aliens: processing of applications applications for *family*..."

Original name: rapport "Office belge de l'immigration: traitement des demandes de regroupement familial"

ID: 2020_02_RegroupementFamilial
Belgian SAI (Cour des comptes), 2020
French, 60 pages, 17,779 words

report 'Full Unemployment Benefits – Prevention and Detection of Undue Payments'

"...Principles of supervision of *family* categories 24 3.2.2

ID: 2018_04_AllocationChomageComple
Belgian SAI (Cour des comptes), 2018
French, 56 pages, 15,807 words





report 'Belgian Immigration Office : Processing Applications for Family Reunification'

Highlighted for query: family reunification

Aliens Office: processing of applications family reunification	Office des étrangers : traitement des demandes de regroupement familial
In 2018, 13,946 visas were issued for family reunification , 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 overstay.	Parmi ceux-ci, 311 visas ont été délivrés sans examen de la demande, en raison d'un dépassement de délai.
In the same year, the Aliens Office received 83,932	Au cours de la même année, l'Office des étrangers a reçu 83.932

Office des étrangers :
traitement des demandes
de regroupement familial

En 2018, 13.946 visas ont été délivrés en vue d'un regroupement familial, soit 43 % des visas de longue durée.

Parmi ceux-ci, 311 visas ont été délivrés sans examen de la demande, en raison d'un dépassement de délai.

Au cours de la même année, l'Office des étrangers a reçu 83.932 demandes de séjour (nouvelles demandes ou demandes de prolongation).

Le regroupement familial est une procédure qui permet aux personnes étrangères dont un membre de la famille séjourne en Belgique de venir le rejoindre à certaines conditions. Ce

Aliens Office:

processing of applications

family reunification

In 2018, 13,946 visas were issued for family reunification, representing 43% of long-term visas.

Of these, 311 visas were issued without examination of the application due to an

Office des étrangers :

traitement des demandes

de regroupement familial

En 2018, 13.946 visas ont été délivrés en vue d'un regroupement familial, soit 43 % des visas de longue durée.

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** 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
 - ...



CLIR feedback questions

1. Would the CLIR tool be useful for you in your work?
2. What kind of tool would be useful for you?

Question Answering

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:**
 - › <http://zilinec.me/question-answering/>
 - › try the default context
 - ask: What is the main expertise of UFAL?
 - ask: Where is the institute located?
 - ask: How many PhD students are there?

Masked Language Models

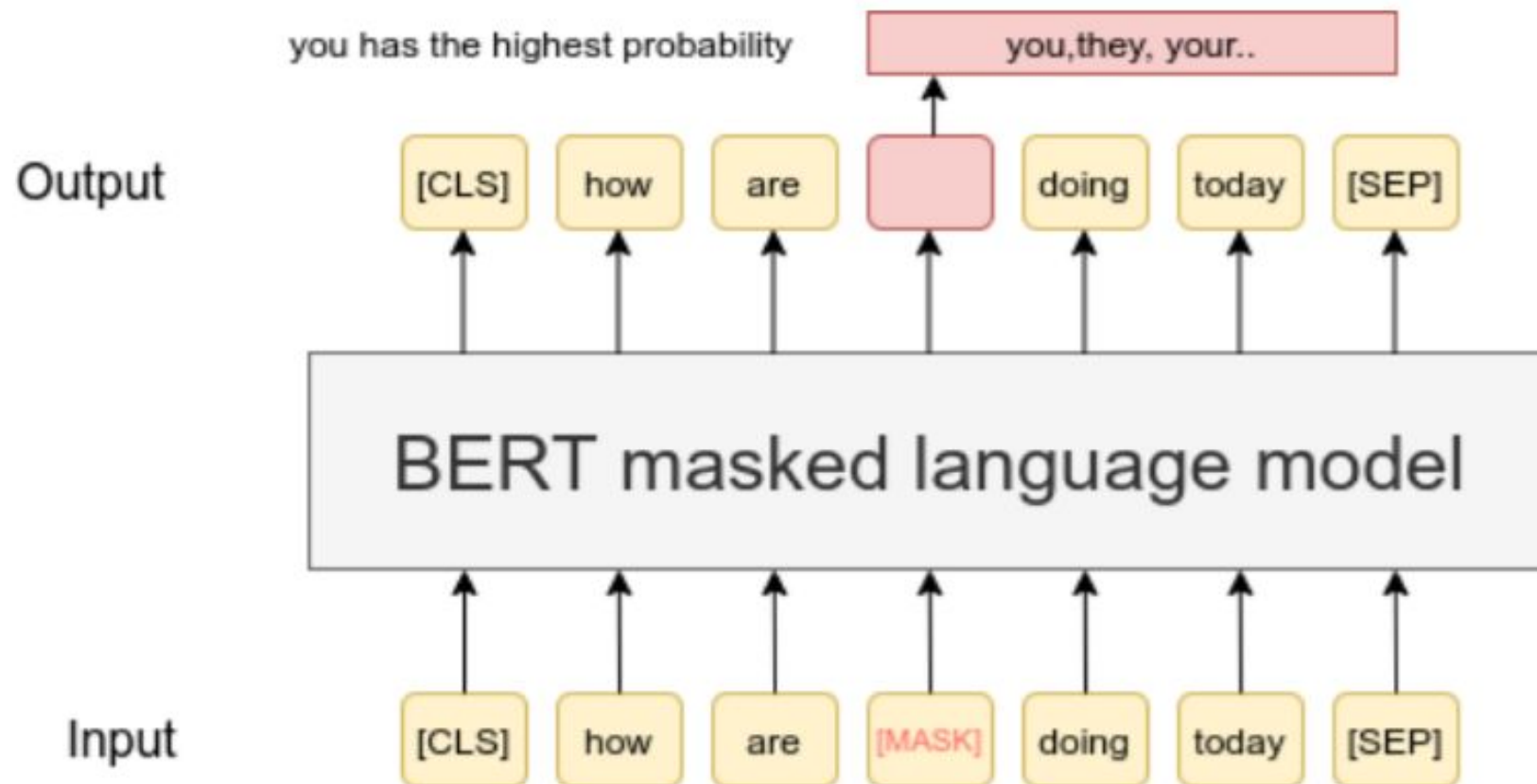
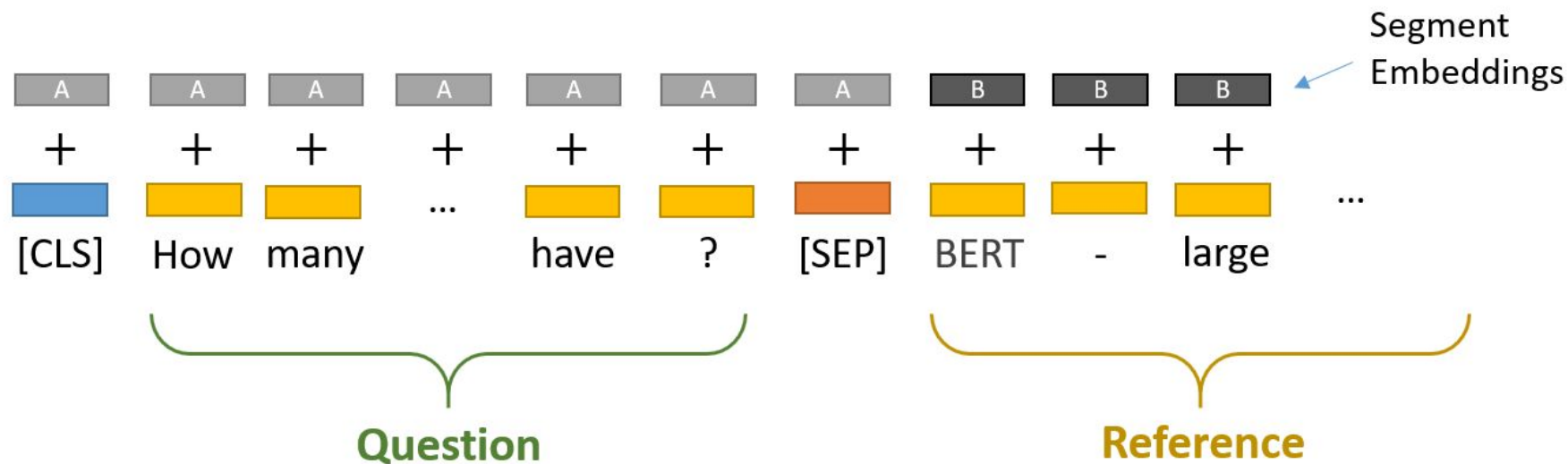


Image source: https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

BERT = Bidirectional Encoder Representations from Transformers

Question Answering with BERT



Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

Question Answering with BERT

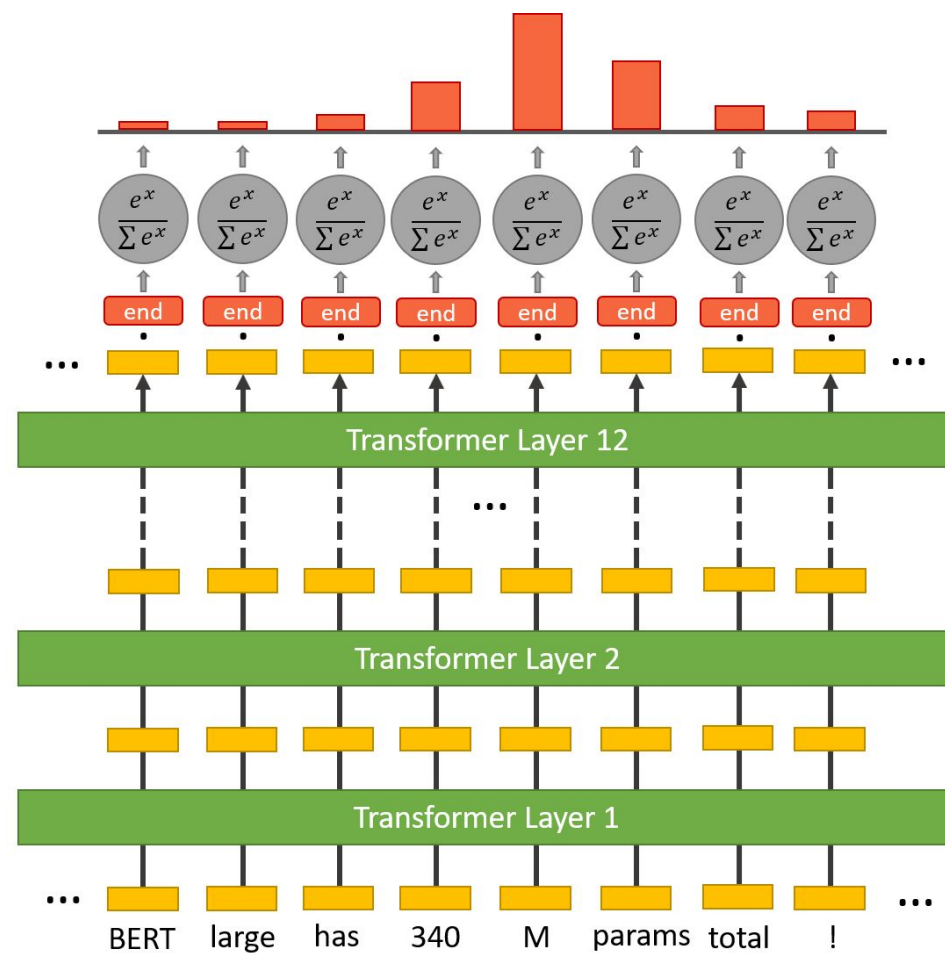
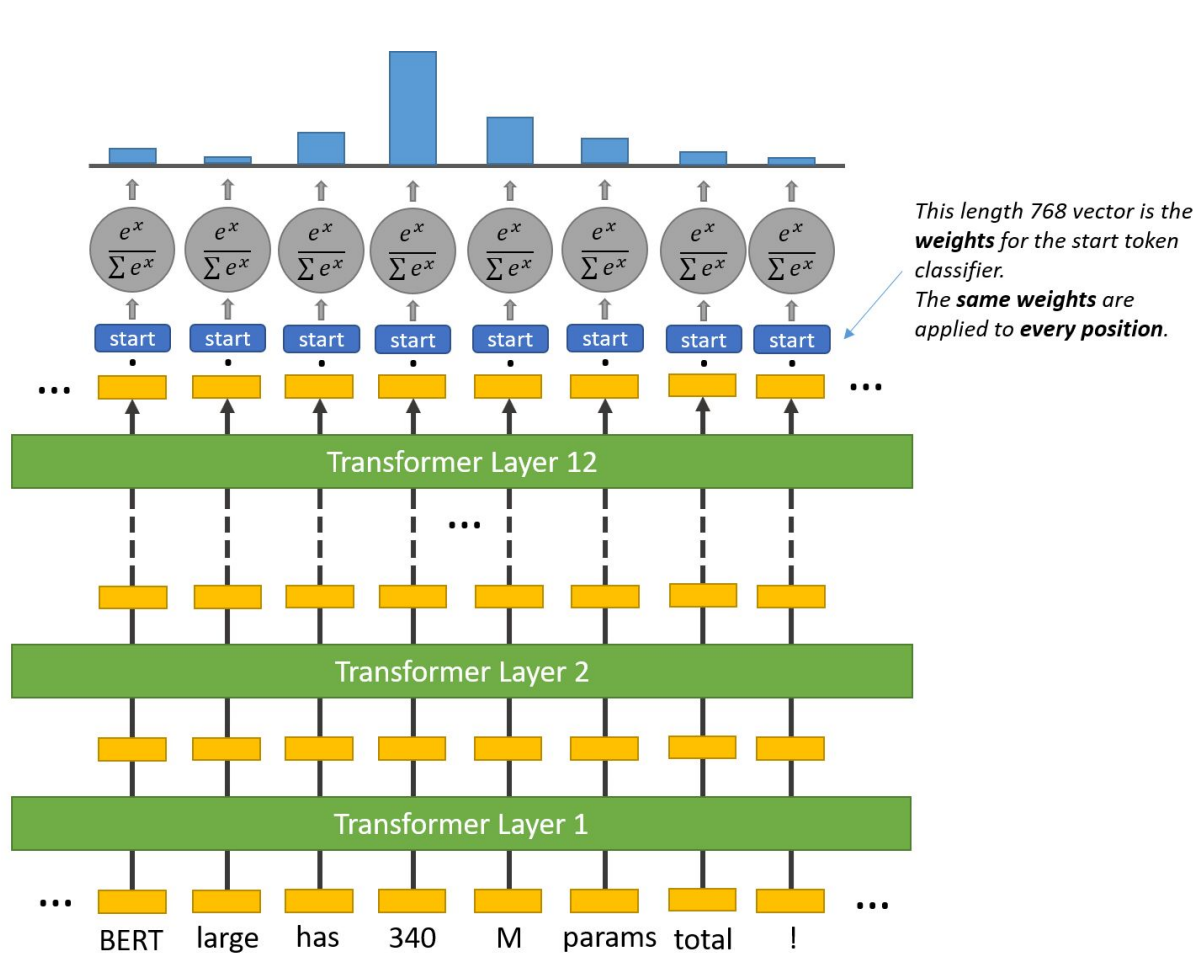


Image source: <https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/>

Question Answering – Task 2

<http://zilinec.me/question-answering/>

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 Kadaň, 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?

Question Answering – Task 3

<http://zilinec.me/question-answering/>

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

Towards Automatic Summarization of Meetings

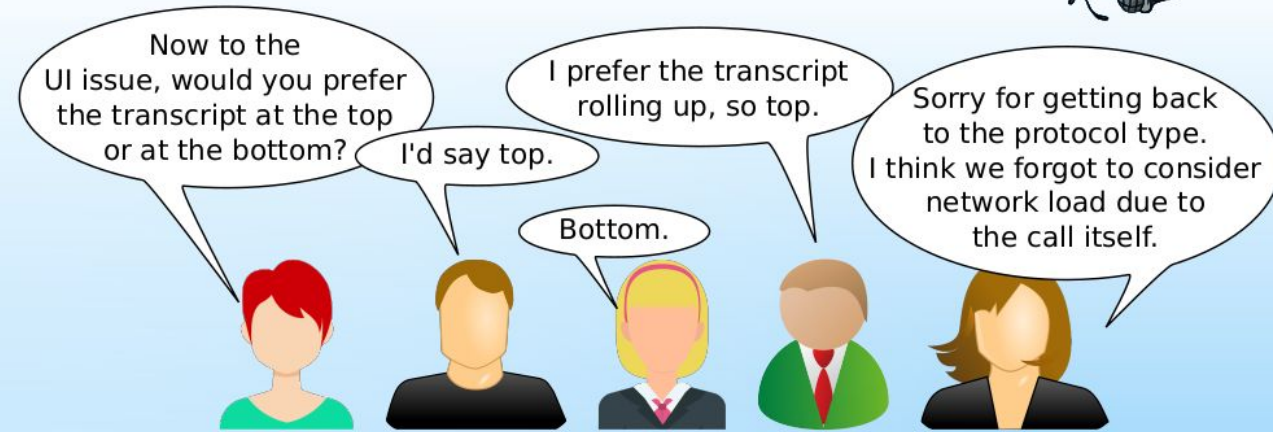
Minuting High-Level Overview



- ▶ Input: A hierarchical agenda.
- ▶ Speech recorded of the fly.
 - ▶ Distinguishing speakers.
- ▶ Full transcript available
 - ▶ Possibly for immediate manual correction.
- ▶ Minuting: Agenda gets populated with summaries of statements.

Possible difference from summarization:

- ▶ We prefer to keep all information, only deduplicate.



Original agenda as prepared by the organizer beforehand:

- Protocol type: push or pull?
- Layout of the user interface:
 - Transcript grows at the top or bottom of the document?
 - Or in a side pane?

Shared document, everyone allowed to edit.

Starts with the agenda and gets populated by Automatic Minuting (AM)

- Protocol type: push or pull?
 - (AM) > Pull easier to implement.
 - (AM) > Updates can get lost with push *in case the user*
 - (AM) > Consider network load.
- Layout of the user interface:
 - Transcript grows at the top or bottom of the document?
 - (AM) > Top (AM) > Bottom (AM) > Top, transcript rolling up.
 - Or in a side pane?

Transcript, optionally editable to correct ASR errors:

- 11:03 Sorry for getting back to the protocol type. I think we forgot ...
- 11:02 I prefer the transcript rolling up, so top.
- 11:02 Bottom
- ...

Sample Output



DATE : 2021-07-21

ATTENDEES : PERSON4, PERSON5, PERSON8,
PERSON10, PERSON13

SUMMARY-

- ▶ The deadline for the project is next Monday, June 15th. Someone from the project needs to be registered there. PERSON8 will try to register today.
- ▶ PERSON13 is going with PERSON4 to LOCATION5. They have a meeting before lunch on Monday. They have one more paper, she wants to submit it to Archive and PROJECT8 so that someone can read it.
- ▶ PERSON10 is on holiday for next two days. They have written one and half paragraph of the book yesterday, and will work on the book from now on.
- ▶ PERSON4 will write half of the chapters.
- ▶ PERSON8 will organize the chapters. They added some information from papers. They will write a preface to the book. He needs to generate, to get the similar metrics from the PROJECT3 and the rest.
- ▶ PERSON5 is going to write his survey. They will work with PERSON8.

- ▶ ALL are working on the papers. The deadline for feedback is at the end of June. The reviewers for PROJECT5 need to be at least a professor, but don't have to be from the university. The grant will be 5000 for it. The deadline for PROJECT7 should be in November. The conference will be virtualized and take place in 2021.
- ▶ PERSON8, PERSON13, PERSON5 and PERSON10 discussed the details of the conference. The abstract submission is on Monday, June 15th. PERSON5 and PERSON8 are going to write a survey for the project. They want to introduce new people to it.
- ▶ ALL discussed about the amount of money they are getting from the university. The money for this year cannot be used for bonuses. PERSON7 bought the computer that he is now using for some grant.
- ▶ PERSON8 got a mail from PR person saying that they can come to the official event.

Further Details on Minuting

...in ELITR deliverables: <https://elitr.eu/deliverables/>

- D1.5: Minuting data, see Section 5.1
- D1.6: Further updates on the data
- D6.5: How the minuting can be put to operation for remote calls; sample outputs

Training data are the **critical limitation**.

- ELITR collected a corpus of ~100 hours of English meetings + ~100 hours of Czech meetings
 - › ELITR provided manual summaries.
- ELITR/ÚFAL is and will be **searching for more such data**.
 - › <https://elitr.eu/recipe-for-miracles-to-happen/>

Possible Ideas for Future Collaboration

Ideas for Brainstorming

Our expertise is mainly in the **sequence-to-sequence** processing, for texts, speech. We also like **implicit structure**: structure in unstructured data.

- Meetings assistance
 - › Live transcription
 - › Live translation
 - › Live summarization
- Decision assistance
 - › Summarization? Summarization into text *and* images?
 - › Automated data analysis, creative visualization?
- Privacy
 - › Can all the processing happen on site?
 - › Can we find a way to process encrypted sound into encrypted transcript?
- Personal data vs. Data sharing
 - › Deep learning critically needs training data.
 - › People are generally reluctant to share the data.
 - › Can we find settings in which they are willing to share (and are not harmed by the sharing)?

Formal Aspects of Collaboration

- EU Research and Innovation Projects (3 yrs, 3+ countries)
 - › We prefer **Research and Innovation Actions (RIA)**.
 - Typically 4-6 partners, 2 of which are companies.
 - 3 EU countries are a minimum, more is better, south is better.
 - Ideally, the goal **comes from the industry, the users**.
 - Need to strike a good balance between blue skies and something practically usable.
 - › Other types are **Innovation Actions (IA)**.
 - More partners, less intense collaboration; more of management.
 - › ...current calls are on “eXtended Reality” and “AI for Human Empowerment”
- Spořitelna as a Partner of Matfyz, Charles University
 - › ?
- Contractual research (“Smluvní výzkum”)
 - › ÚFAL can carry out research work (or provide/create specific data) for money.
 - › Generally not our main interest.
- Random support:
 - › E.g. catering for the hackathon on MT for Ukrainian

Discussion

Questionnaire - evaluation

Thank you for your attention



Thank you for your attention

Charles University, Czech Republic

Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

ÚFAL

bojar@ufal.mff.cuni.cz
kapralova@ufal.mff.cuni.cz