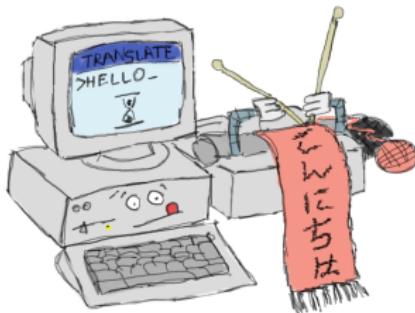


# Machine translation and artificial intelligence

Mgr. Martin Popel, Ph.D.

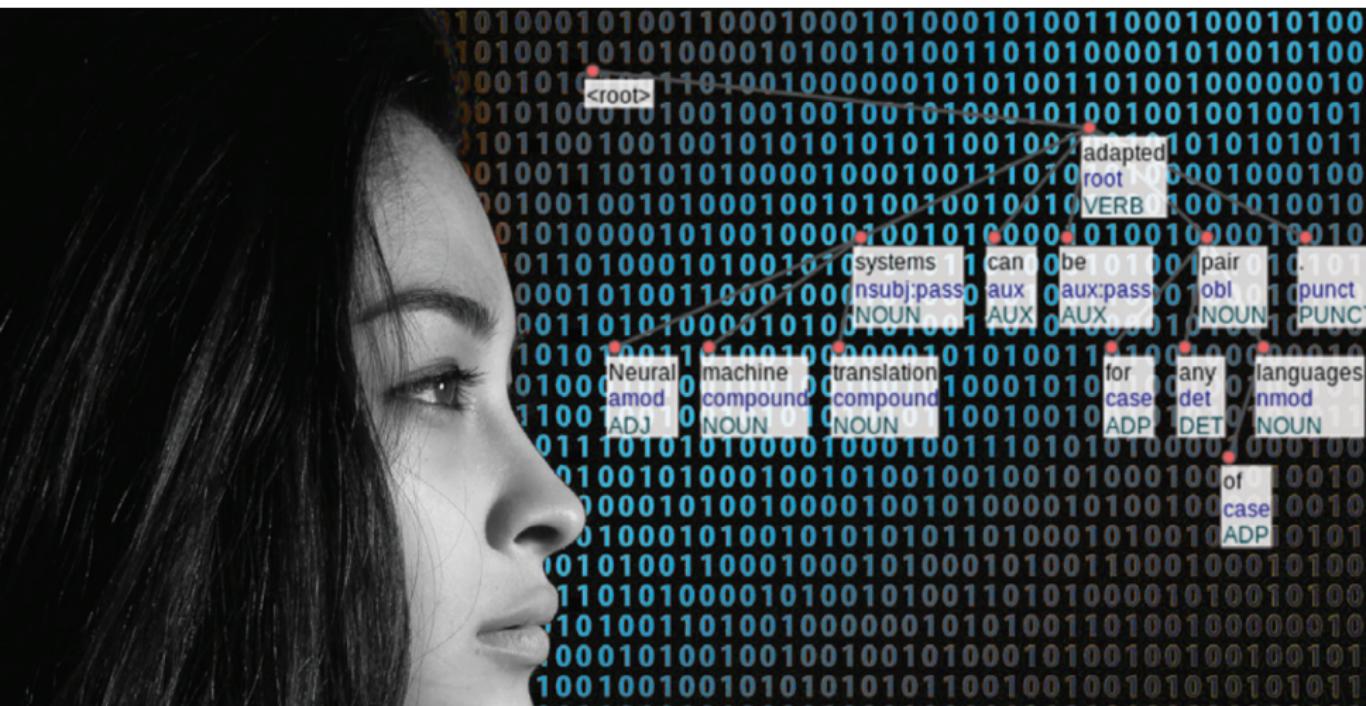
ÚFAL (Institute of Formal and Applied Linguistics), MFF UK

2021-11-09, Physics Institute seminar



source	Great talkers are little doers.
Yandex	Velké talkers jsou trochu činitelé.
Bing	Velcí vysílačky jsou malí činitelé.
Google	Velcí mluvčí jsou malí lidé.
TectoMT	Velcí řečníci jsou malí vrazi.
CUBBITT	Velcí mluvkové jsou malí dřiči.

## Natural Language Processing?



## Spell check vs. grammar check

3

Chlapci šly.  
Chlapec šli do školy.

## Spell check vs. grammar check

3

Dívce nešly hodinky. Chlapci šly.  
Chlapec šli do školy.

# Spell check vs. grammar check

3

Dívce nešly hodinky. Chlapci šly.  
Kdo kam co donesl? Chlapec šli do školy.



# Spell check vs. grammar check

3

<http://ufal.cz/korektor>

Korektor-sample.rtf — Edited

Potkávám je na každém **krou**.

Did you mean kroku?

Lední medvěd byl schován za **krou**.

Vláďa nás **přivíral** s otevřenou **náruči**.

Did you mean přivítal? Did you mean náručí?

Kocour **přivíral** oči slastí, když pozoroval kanárka v **klexi**.

Tento víkend raději zůstaneme v **praze**.

Error in capitalization? Did you mean Praze?

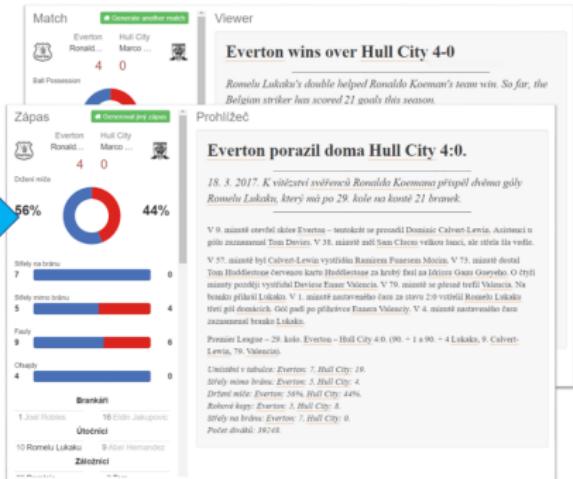
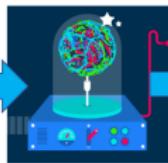
## Automatic news generation

4

more news (FB, Twitter), less journalist, less profit  
need for personalized and instant news

Index	Name	Age	Gender	Address	City	State	Country
0	John Doe	30	Male	123 Main St	New York	NY	USA
1	Jane Smith	25	Female	456 Elm St	Boston	MA	USA
2	Bob Johnson	40	Male	789 Oak St	Chicago	IL	USA
3	Sarah Davis	28	Female	567 Pine St	Los Angeles	CA	USA
4	David Wilson	35	Male	890 Cedar St	Houston	TX	USA
5	Emily Green	22	Female	234 Birch St	Phoenix	AZ	USA
6	Michael Brown	42	Male	654 Maple St	Seattle	WA	USA
7	Karen Taylor	38	Female	987 Cherry St	Tampa	FL	USA
8	James White	20	Male	456 Pine St	Portland	OR	USA
9	Laura Lee	26	Female	789 Cedar St	San Jose	CA	USA
10	Mark Williams	32	Male	567 Birch St	Austin	TX	USA
11	Sarah Thompson	24	Female	234 Maple St	Philadelphia	PA	USA
12	David Wilson	35	Male	890 Pine St	Washington	DC	USA
13	Emily Green	22	Female	987 Cedar St	Baltimore	MD	USA
14	Michael Brown	42	Male	456 Birch St	Nashville	TN	USA
15	Karen Taylor	38	Female	789 Maple St	Atlanta	GA	USA
16	James White	20	Male	567 Pine St	Phoenix	AZ	USA
17	Laura Lee	26	Female	234 Cedar St	Seattle	WA	USA
18	Mark Williams	32	Male	890 Birch St	Tampa	FL	USA
19	Sarah Thompson	24	Female	567 Maple St	Portland	OR	USA
20	David Wilson	35	Male	987 Pine St	San Jose	CA	USA
21	Emily Green	22	Female	456 Cedar St	Austin	TX	USA
22	Michael Brown	42	Male	789 Birch St	Philadelphia	PA	USA
23	Karen Taylor	38	Female	567 Maple St	Baltimore	MD	USA
24	James White	20	Male	234 Cedar St	Nashville	TN	USA
25	Laura Lee	26	Female	890 Birch St	Atlanta	GA	USA
26	Mark Williams	32	Male	567 Maple St	Tampa	FL	USA
27	Sarah Thompson	24	Female	987 Cedar St	Phoenix	AZ	USA
28	David Wilson	35	Male	456 Birch St	Seattle	WA	USA
29	Emily Green	22	Female	789 Maple St	Portland	OR	USA
30	Michael Brown	42	Male	567 Pine St	San Jose	CA	USA
31	Karen Taylor	38	Female	234 Cedar St	Austin	TX	USA
32	James White	20	Male	890 Maple St	Philadelphia	PA	USA
33	Laura Lee	26	Female	567 Birch St	Baltimore	MD	USA
34	Mark Williams	32	Male	987 Pine St	Nashville	TN	USA
35	Sarah Thompson	24	Female	456 Cedar St	Atlanta	GA	USA
36	David Wilson	35	Male	789 Maple St	Tampa	FL	USA
37	Emily Green	22	Female	567 Birch St	Phoenix	AZ	USA
38	Michael Brown	42	Male	987 Maple St	Seattle	WA	USA
39	Karen Taylor	38	Female	456 Cedar St	Portland	OR	USA
40	James White	20	Male	789 Pine St	San Jose	CA	USA
41	Laura Lee	26	Female	567 Cedar St	Austin	TX	USA
42	Mark Williams	32	Male	987 Maple St	Philadelphia	PA	USA
43	Sarah Thompson	24	Female	456 Birch St	Baltimore	MD	USA
44	David Wilson	35	Male	789 Cedar St	Nashville	TN	USA
45	Emily Green	22	Female	567 Maple St	Atlanta	GA	USA
46	Michael Brown	42	Male	234 Birch St	Tampa	FL	USA
47	Karen Taylor	38	Female	890 Cedar St	Phoenix	AZ	USA
48	James White	20	Male	567 Maple St	Seattle	WA	USA
49	Laura Lee	26	Female	987 Birch St	Portland	OR	USA
50	Mark Williams	32	Male	456 Cedar St	San Jose	CA	USA
51	Sarah Thompson	24	Female	789 Maple St	Austin	TX	USA
52	David Wilson	35	Male	567 Birch St	Philadelphia	PA	USA
53	Emily Green	22	Female	987 Cedar St	Baltimore	MD	USA
54	Michael Brown	42	Male	456 Maple St	Nashville	TN	USA
55	Karen Taylor	38	Female	234 Cedar St	Atlanta	GA	USA
56	James White	20	Male	890 Maple St	Tampa	FL	USA
57	Laura Lee	26	Female	567 Birch St	Phoenix	AZ	USA
58	Mark Williams	32	Male	987 Cedar St	Seattle	WA	USA
59	Sarah Thompson	24	Female	456 Maple St	Portland	OR	USA
60	David Wilson	35	Male	789 Birch St	San Jose	CA	USA
61	Emily Green	22	Female	567 Cedar St	Austin	TX	USA
62	Michael Brown	42	Male	987 Maple St	Philadelphia	PA	USA
63	Karen Taylor	38	Female	456 Birch St	Baltimore	MD	USA
64	James White	20	Male	789 Cedar St	Nashville	TN	USA
65	Laura Lee	26	Female	567 Maple St	Atlanta	GA	USA
66	Mark Williams	32	Male	234 Cedar St	Tampa	FL	USA
67	Sarah Thompson	24	Female	890 Birch St	Phoenix	AZ	USA
68	David Wilson	35	Male	567 Cedar St	Seattle	WA	USA
69	Emily Green	22	Female	987 Maple St	Portland	OR	USA
70	Michael Brown	42	Male	456 Birch St	San Jose	CA	USA
71	Karen Taylor	38	Female	789 Cedar St	Austin	TX	USA
72	James White	20	Male	567 Maple St	Philadelphia	PA	USA
73	Laura Lee	26	Female	987 Birch St	Baltimore	MD	USA
74	Mark Williams	32	Male	456 Cedar St	Nashville	TN	USA
75	Sarah Thompson	24	Female	234 Maple St	Atlanta	GA	USA
76	David Wilson	35	Male	890 Cedar St	Tampa	FL	USA
77	Emily Green	22	Female	567 Birch St	Phoenix	AZ	USA
78	Michael Brown	42	Male	987 Maple St	Seattle	WA	USA
79	Karen Taylor	38	Female	456 Cedar St	Portland	OR	USA
80	James White	20	Male	789 Maple St	San Jose	CA	USA
81	Laura Lee	26	Female	567 Birch St	Austin	TX	USA
82	Mark Williams	32	Male	987 Cedar St	Philadelphia	PA	USA
83	Sarah Thompson	24	Female	456 Cedar St	Baltimore	MD	USA
84	David Wilson	35	Male	789 Maple St	Nashville	TN	USA
85	Emily Green	22	Female	567 Birch St	Atlanta	GA	USA
86	Michael Brown	42	Male	234 Cedar St	Tampa	FL	USA
87	Karen Taylor	38	Female	890 Maple St	Phoenix	AZ	USA
88	James White	20	Male	567 Cedar St	Seattle	WA	USA
89	Laura Lee	26	Female	987 Maple St	Portland	OR	USA
90	Mark Williams	32	Male	456 Birch St	San Jose	CA	USA
91	Sarah Thompson	24	Female	789 Cedar St	Austin	TX	USA
92	David Wilson	35	Male	567 Cedar St	Philadelphia	PA	USA
93	Emily Green	22	Female	987 Maple St	Baltimore	MD	USA
94	Michael Brown	42	Male	456 Birch St	Nashville	TN	USA
95	Karen Taylor	38	Female	234 Cedar St	Atlanta	GA	USA
96	James White	20	Male	890 Cedar St	Tampa	FL	USA
97	Laura Lee	26	Female	567 Birch St	Phoenix	AZ	USA
98	Mark Williams	32	Male	987 Maple St	Seattle	WA	USA
99	Sarah Thompson	24	Female	456 Cedar St	Portland	OR	USA
100	David Wilson	35	Male	789 Cedar St	San Jose	CA	USA
101	Emily Green	22	Female	567 Maple St	Austin	TX	USA
102	Michael Brown	42	Male	234 Cedar St	Philadelphia	PA	USA
103	Karen Taylor	38	Female	890 Maple St	Baltimore	MD	USA
104	James White	20	Male	567 Cedar St	Nashville	TN	USA
105	Laura Lee	26	Female	987 Cedar St	Atlanta	GA	USA
106	Mark Williams	32	Male	456 Maple St	Tampa	FL	USA
107	Sarah Thompson	24	Female	789 Cedar St	Phoenix	AZ	USA
108	David Wilson	35	Male	567 Cedar St	Seattle	WA	USA
109	Emily Green	22	Female	987 Maple St	Portland	OR	USA
110	Michael Brown	42	Male	456 Cedar St	San Jose	CA	USA
111	Karen Taylor	38	Female	234 Cedar St	Austin	TX	USA
112	James White	20	Male	890 Cedar St	Philadelphia	PA	USA
113	Laura Lee	26	Female	567 Cedar St	Baltimore	MD	USA
114	Mark Williams	32	Male	987 Maple St	Nashville	TN	USA
115	Sarah Thompson	24	Female	456 Cedar St	Atlanta	GA	USA
116	David Wilson	35	Male	789 Cedar St	Tampa	FL	USA
117	Emily Green	22	Female	567 Cedar St	Phoenix	AZ	USA
118	Michael Brown	42	Male	987 Cedar St	Seattle	WA	USA
119	Karen Taylor	38	Female	456 Cedar St	Portland	OR	USA
120	James White	20	Male	789 Maple St	San Jose	CA	USA
121	Laura Lee	26	Female	567 Cedar St	Austin	TX	USA
122	Mark Williams	32	Male	234 Cedar St	Philadelphia	PA	USA
123	Sarah Thompson	24	Female	890 Cedar St	Baltimore	MD	USA
124	David Wilson	35	Male	567 Cedar St	Nashville	TN	USA
125	Emily Green	22	Female	987 Cedar St	Atlanta	GA	USA
126	Michael Brown	42	Male	456 Cedar St	Tampa	FL	USA
127	Karen Taylor	38	Female	234 Cedar St	Phoenix	AZ	USA
128	James White	20	Male	890 Cedar St	Seattle	WA	USA
129	Laura Lee	26	Female	567 Cedar St	Portland	OR	USA
130	Mark Williams	32	Male	987 Cedar St	San Jose	CA	USA
131	Sarah Thompson	24	Female	456 Cedar St	Austin	TX	USA
132	David Wilson	35	Male	789 Cedar St	Philadelphia	PA	USA
133	Emily Green	22	Female	567 Cedar St	Baltimore	MD	USA
134	Michael Brown	42	Male	987 Cedar St	Nashville	TN	USA
135	Karen Taylor	38	Female	456 Cedar St	Atlanta	GA	USA
136	James White	20	Male	789 Cedar St	Tampa	FL	USA
137	Laura Lee	26	Female	567 Cedar St	Phoenix	AZ	USA
138	Mark Williams	32	Male	987 Cedar St	Seattle	WA	USA
139	Sarah Thompson	24	Female	456 Cedar St	Portland	OR	USA
140	David Wilson	35	Male	789 Cedar St	San Jose	CA	USA
141	Emily Green	22	Female	567 Cedar St	Austin	TX	USA
142	Michael Brown	42	Male	234 Cedar St	Philadelphia	PA	USA
143	Karen Taylor	38	Female	890 Cedar St	Baltimore	MD	USA
144	James White	20	Male	567 Cedar St	Nashville	TN	USA
145	Laura Lee	26	Female	987 Cedar St	Atlanta	GA	USA
146	Mark Williams	32	Male	456 Cedar St	Tampa	FL	USA
147	Sarah Thompson	24	Female	789 Cedar St	Phoenix	AZ	USA
148	David Wilson	35	Male	567 Cedar St	Seattle	WA	USA
149	Emily Green	22	Female	987 Cedar St	Portland	OR	USA
150	Michael Brown	42	Male	456 Cedar St	San Jose	CA	USA
151	Karen Taylor	38	Female	234 Cedar St	Austin	TX	USA
152	James White	20	Male	890 Cedar St	Philadelphia	PA	USA
153	Laura Lee	26	Female	567 Cedar St	Baltimore	MD	USA
154	Mark Williams	32	Male	987 Cedar St	Nashville	TN	USA
155	Sarah Thompson	24	Female	456 Cedar St	Atlanta	GA	USA
156	David Wilson	35	Male	789 Cedar St	Tampa	FL	USA
157	Emily Green	22	Female	567 Cedar St	Phoenix	AZ	USA
158	Michael Brown	42	Male	987 Cedar St	Seattle	WA	USA
159	Karen Taylor	38	Female	456 Cedar St	Portland	OR	USA
160	James White	20	Male	789 Cedar St	San Jose	CA	USA
161	Laura Lee	26	Female	567 Cedar St	Austin	TX	USA
162	Mark Williams	32	Male	234 Cedar St	Philadelphia	PA	USA
163	Sarah Thompson	24	Female	890 Cedar St	Baltimore	MD	USA
164	David Wilson	35	Male	567 Cedar St	Nashville	TN	USA
165	Emily Green	22	Female	987 Cedar St	Atlanta	GA	USA
166	Michael Brown	42	Male	456 Cedar St	Tampa	FL	USA
167	Karen Taylor	38	Female	234 Cedar St	Phoenix	AZ	USA
168	James White	20	Male	890 Cedar St	Seattle	WA	USA
169	Laura Lee	26	Female	567 Cedar St	Portland	OR	USA
170	Mark Williams	32	Male	987 Cedar St	San Jose	CA	USA
171	Sarah Thompson	24	Female	456 Cedar St	Austin	TX	USA
172	David Wilson	35	Male	789 Cedar St	Philadelphia	PA	USA
173	Emily Green	22	Female	567 Cedar St	Baltimore	MD	USA
174	Michael Brown	42	Male	987 Cedar St	Nashville	TN	USA
175	Karen Taylor	38	Female	456 Cedar St	Atlanta	GA	USA
176	James White	20	Male	789 Cedar St	Tampa	FL	USA
177	Laura Lee	26	Female	567 Cedar St	Phoenix	AZ	USA
178	Mark Williams	32	Male	987 Cedar St	Seattle	WA	USA
179	Sarah Thompson	24	Female	456 Cedar St	Portland	OR	USA
180	David Wilson	35	Male	789 Cedar St	San Jose	CA	USA
181	Emily Green	22	Female	567 Cedar St	Austin	TX	USA
182	Michael Brown	42	Male	234 Cedar St	Philadelphia	PA	USA
183	Karen Taylor	38	Female	890 Cedar St	Baltimore	MD	USA
184	James White	20	Male	567 Cedar St	Nashville	TN	USA
185	Laura Lee	26	Female	987 Cedar St	Atlanta	GA	USA
186	Mark Williams	32	Male	456 Cedar St	Tampa	FL	USA
187	Sarah Thompson	24	Female	789 Cedar St	Phoenix	AZ	USA
188	David Wilson	35	Male	567 Cedar St	Seattle	WA	USA
189	Emily Green	22	Female	987 Cedar St	Portland	OR	USA
190	Michael Brown	42	Male	456 Cedar St	San Jose	CA	USA
191	Karen Taylor	38	Female	234 Cedar St	Austin	TX	USA
192	James White	20	Male	890 Cedar St	Philadelphia	PA	USA
193	Laura Lee	26	Female	567 Cedar St	Baltimore	MD	USA
194	Mark Williams	32	Male	987 Cedar St	Nashville	TN	USA
195	Sarah Thompson	24	Female	456 Cedar St	Atlanta	GA	USA
196	David Wilson	35	Male	789 Cedar St	Tampa	FL	USA
197	Emily Green	22	Female	567 Cedar St	Phoenix	AZ	USA
198	Michael Brown	42	Male	987 Cedar St	Seattle	WA	USA
199	Karen Taylor	38	Female	456 Cedar St	Portland	OR	USA
200	James White	20	Male	789 Cedar St	San Jose	CA	USA

Structured data  
(real-time & historical)



## Article / Report

Byl by to rytíř, kde v pláně hřích vzlet,  
Vědě jsem jse seheldo na přídoutně v světě si nezastavá:  
„Ukryjemné, chvěla, milý nás jest

Kolem jsou jest vyhrávaných  
A svítí co pláčem, rád pravil:  
Ale plná jízdo zaporodilo se, vys.

již dávno vás poháru a vlanných rány,  
v jablonění je píše je i v kristování,

srdce v své ženských svém  
v obly pětky tam a vzíti,  
na kóňku je, milý svěžek.

I'll come a bit later on my own.

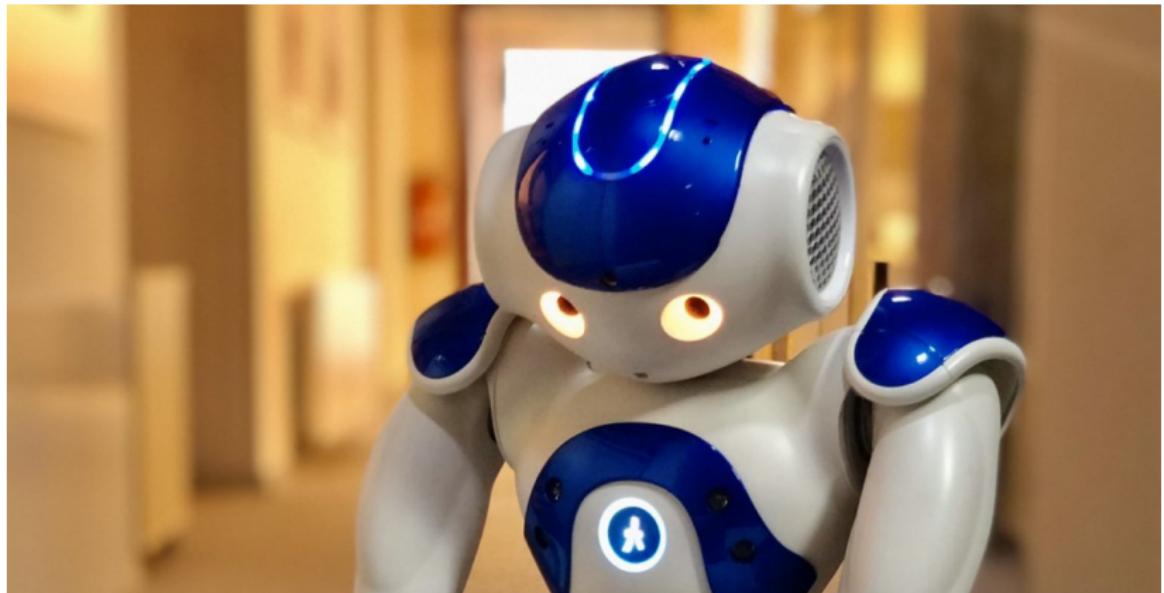
I'll come a bit later on my own.  
Sem čelist ještě na své milé.

# Can a robot write a theatre play?

6

MFF + DAMU + Švanda theatre

premiere in February 2021 (100 years after R.U.R.)



<https://www.theaitre.com>

# Korupci často prozradí „kapříci“

Vyšetřovací tým právníků a forenzních analytiků hledá ve firmách důkazy o korupci. Proloží kódovanou řeč i šifrovací aplikace

KATEŘINA KOLÁŘOVÁ

**V**e druhém patře moderní pražské kancelářské budovy Nile House usedá k jednacímu stolu pestro složený tým odborníků. Právníci, forenzní analytičtí a vyšetřovatelé zjistují, jestli byly v prošetrování společnosti globálního významu uzavřeny pro firmu nevhodné smlouvy. Experti společnosti Deloitte zkoumají, jestli zaměstnanci vyšetřované společnosti „šíří na rukou“ dodavatelů služeb, a za uplatek mu umožnili vyhrát lukrativní zakázku. K vyšetřování používají speciální techniky automatické analýzy dat, která prohlédne i kódovanou řeč.

„Velmí často řešíme právě vztahy dodavatelů s nákupním oddělením, ty jsou problematické témař v každé prošetrované společnosti,“ vysvětluje Jaroslava Kračúnová, advokátka a partnerka kanceláře Ambur & Dark Deloitte Legal, jež spolupracuje s forenzním týmem vyšetřovatelů. Multidisciplinární tým využívá i podezření z financování terorismu v zahraničí. „Takto závažné trestné činy se velmi těžko prokazují izolovaně v prošetrování společnosti klienta. Proto často, paralelně s naším řešením, probíhá i policejní vyšetřování,“ říká Kračúnová.

## Rychlý zásah



**Multidisciplinární tým.** Vyšetřování podezření z korupce nebo projevů sexuálního hrazení na pracovišti spojuje práci rozdílných vědních oborů. Analytický tým vede absolventka matematickovo-fyzikální fakulty Kateřina Veselovská (vlevo), právník-advokátka Jaroslava Kračúnová (vpravo).

dostatečně odůvodněné. Věc řešíme po stránce pracovněprávní, a s ohledem na ochranu osobních údajů a ochranu soukromí,“ popisuje Kračúnová začátek vyšetřování. Po dokončení právních kroků

ké analýzy, kterou v Deloitte využívají téměř dva roky, a to ve 28 jazyčích.

## Rozpoznaní kódované řeči

„Na základě blízkosti slovní zásoby

Důkazem manipulace s výsledky tendru je typicky vznik

mohla existovat, protože o vítězi tendru ještě nebylo nákupním oddělením oficiálně rozhodnuto,“ popisuje Kračúnová. Pak už analytici hledají v počítačích konkrétní texury soubor.

**Kateřina Veselovská**

■ Manažerka oddělení Data Analytics Deloitte, kde vede tým pro analýzu nestrukturovaných dat. Absolvovala doktorandský program na Matematickovo-fyzikální fakultě Univerzity Karlovy v Praze. Věnovala se vývoji softwaru pro textovou analytiku a business poradenství v projektech týkajících se oblasti nestrukturovaných a velkých dat. Nyní se zaměřuje zejména na projekty z oboru forenzní analytiky a řezení rizik.

**Jaroslava Kračúnová**

■ Partnerka a advokátka v Deloitte Legal, vede tým Business Integrity. Promována na Právnické fakultě Univerzity Karlovy v Praze. Studovala i právo a management v Innsbrucku. Specializuje se na odsoulovaní hospodářské kriminality, trestní odpovědnost právnických osob, corporate governance a ochranu osobních údajů.

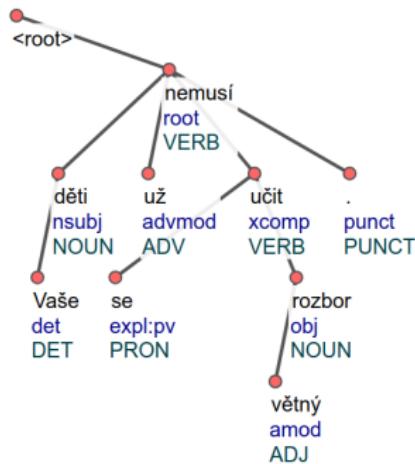
rozvolněné hranice toho, co je v této oblasti už za hranou a co je ještě v pořádku. V některých zemích je to ale něco naprostě nepřirozeného,“ vzpomíná Kračúnová. Byť tím specialistů prokázal, že se

# Automatic syntactic analysis

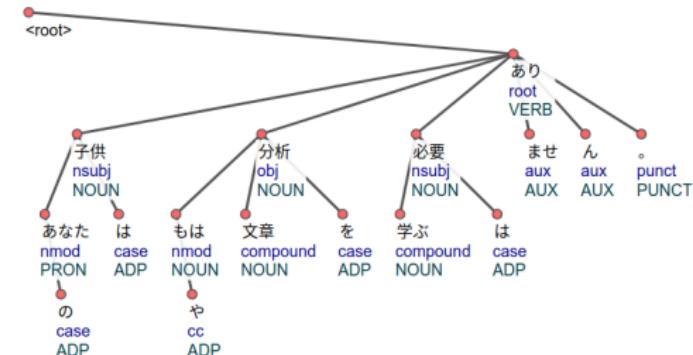
8

aka parsing, available for 50+ languages  
accuracy for Czech about 90% (85% including morphology)

Vaše děti se už nemusí učit větný rozbor.



あなたの子供はもはや文章分析を学ぶ必要はありません。

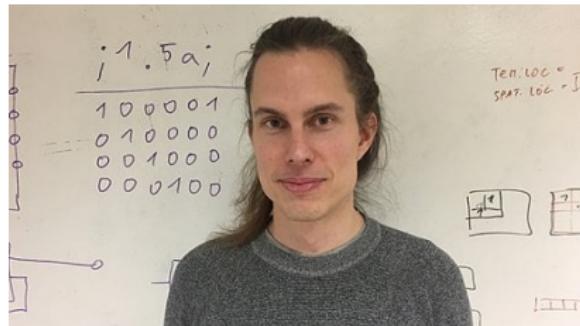


Can you add and subtract numbers?  
What about words and images?

king - man + woman = ?

$$\text{king} - \text{man} + \text{woman} = \text{queen}$$

Tomáš Mikolov, 2012, word2vec

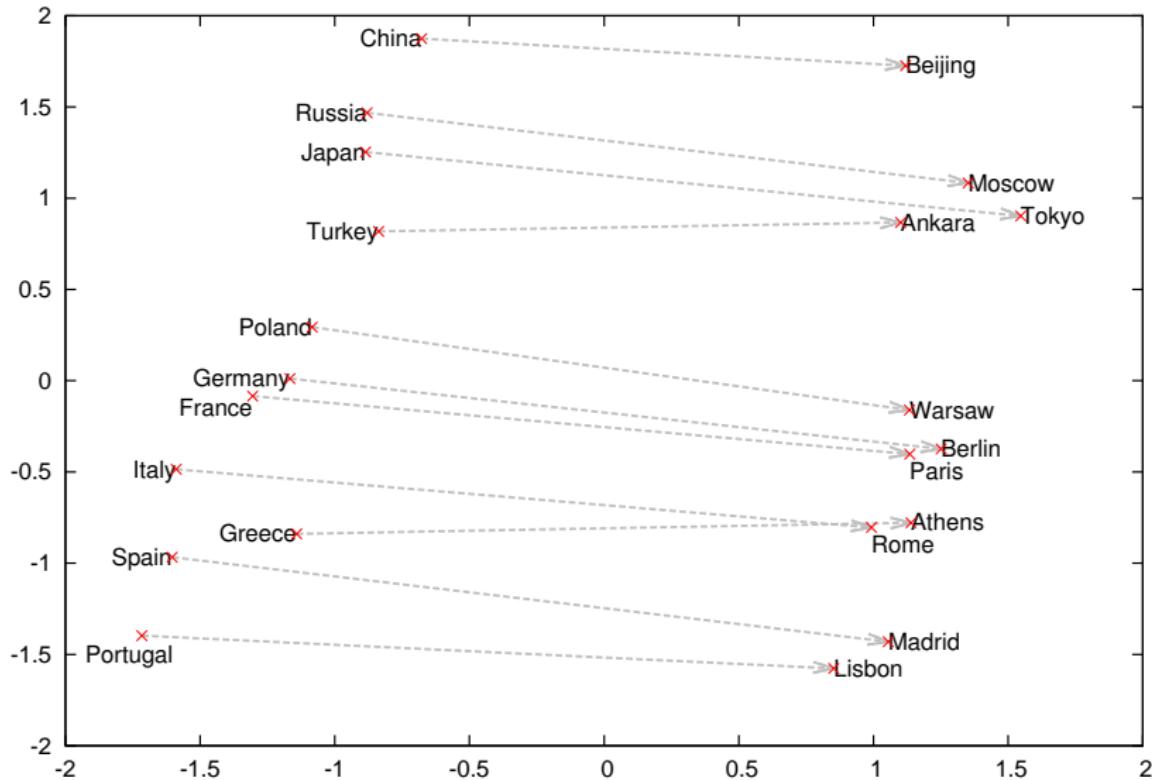


<https://projector.tensorflow.org/>

# Word embeddings

9

Country and Capital Vectors Projected by PCA



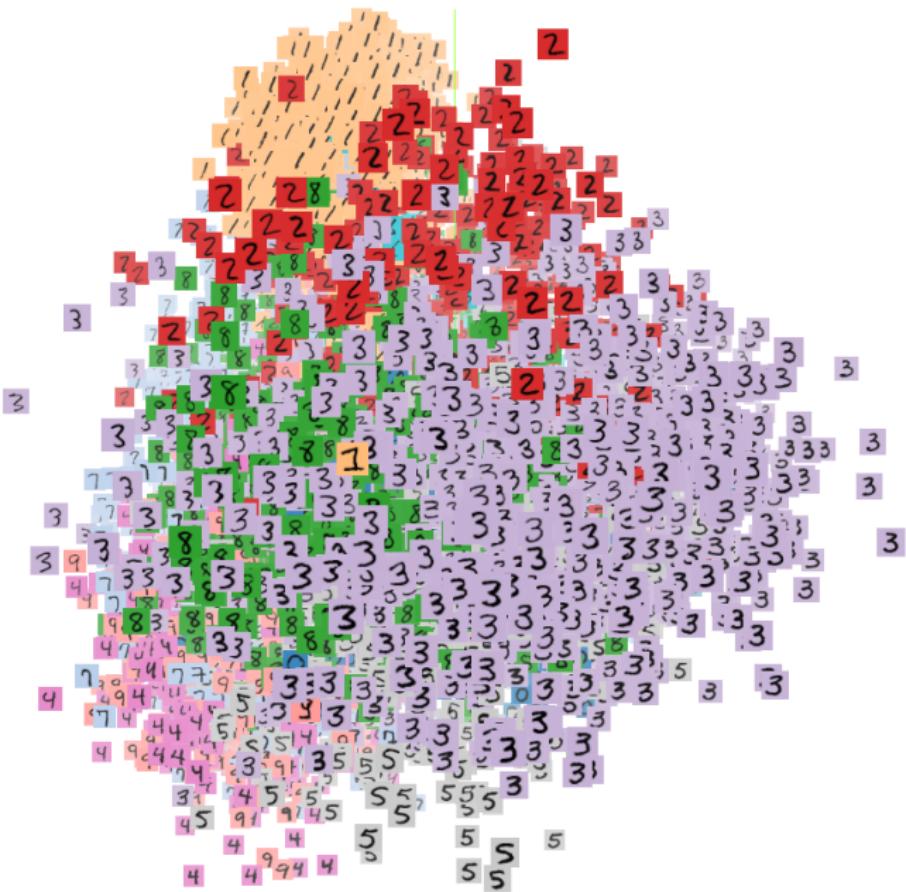
# Word embeddings

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

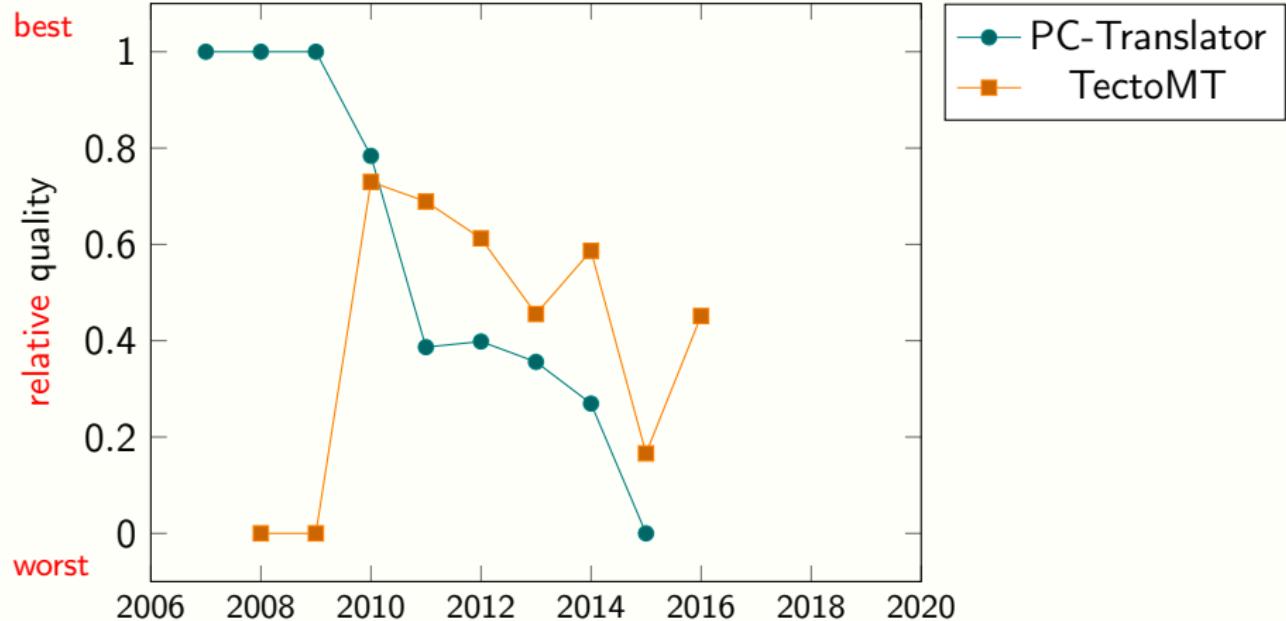
Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

# Word embeddings



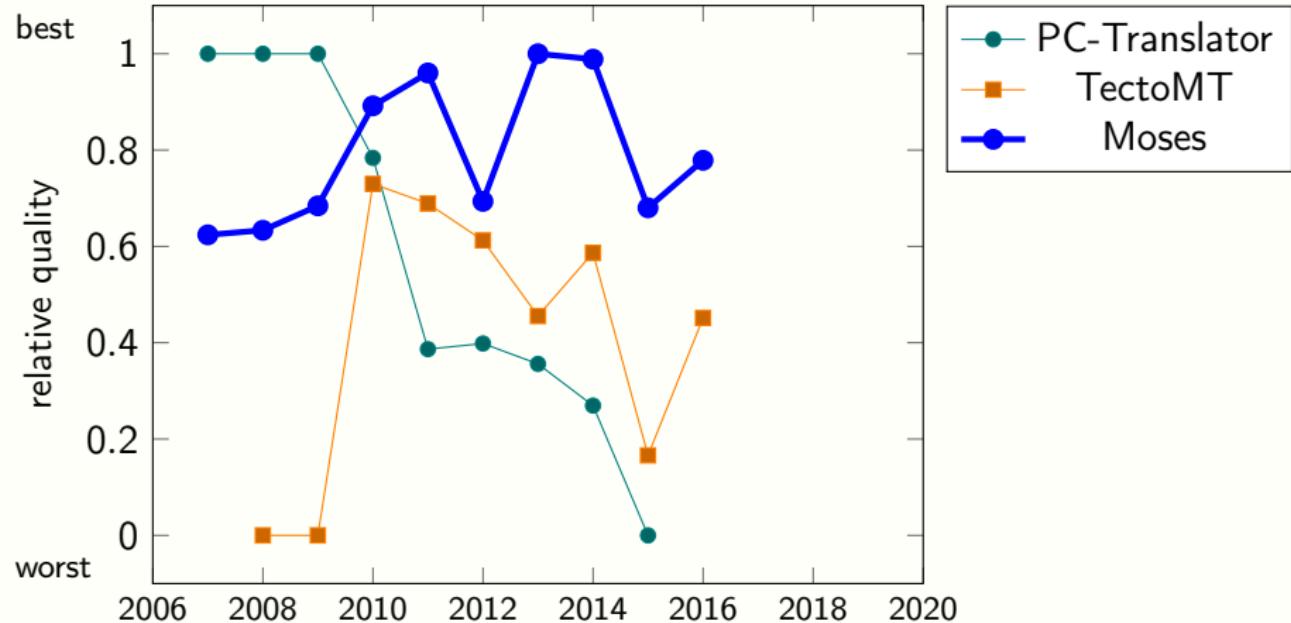
# English→Czech MT in 2007–2020

10



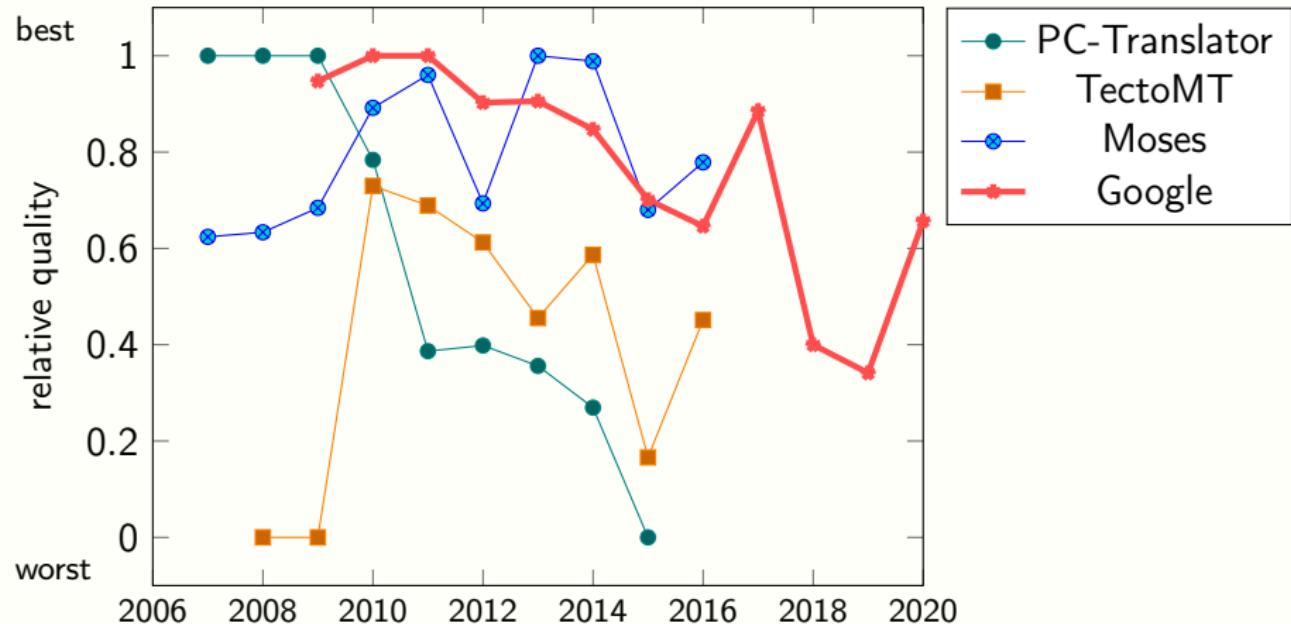
# English→Czech MT in 2007–2020

10



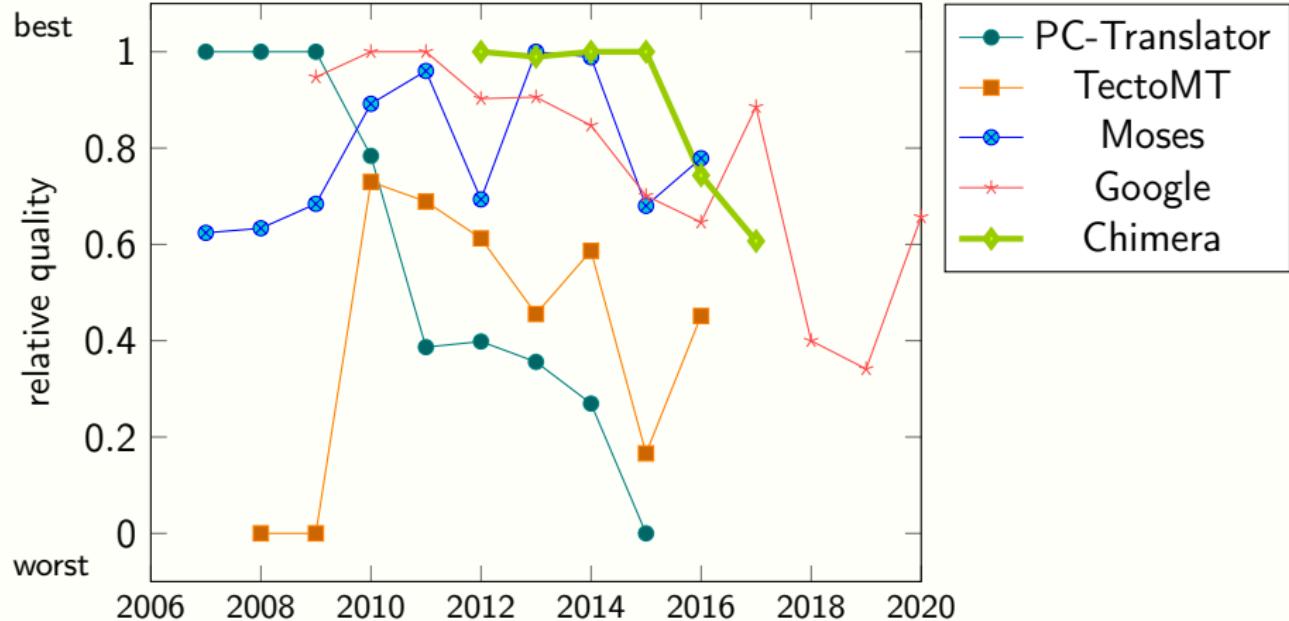
# English→Czech MT in 2007–2020

10



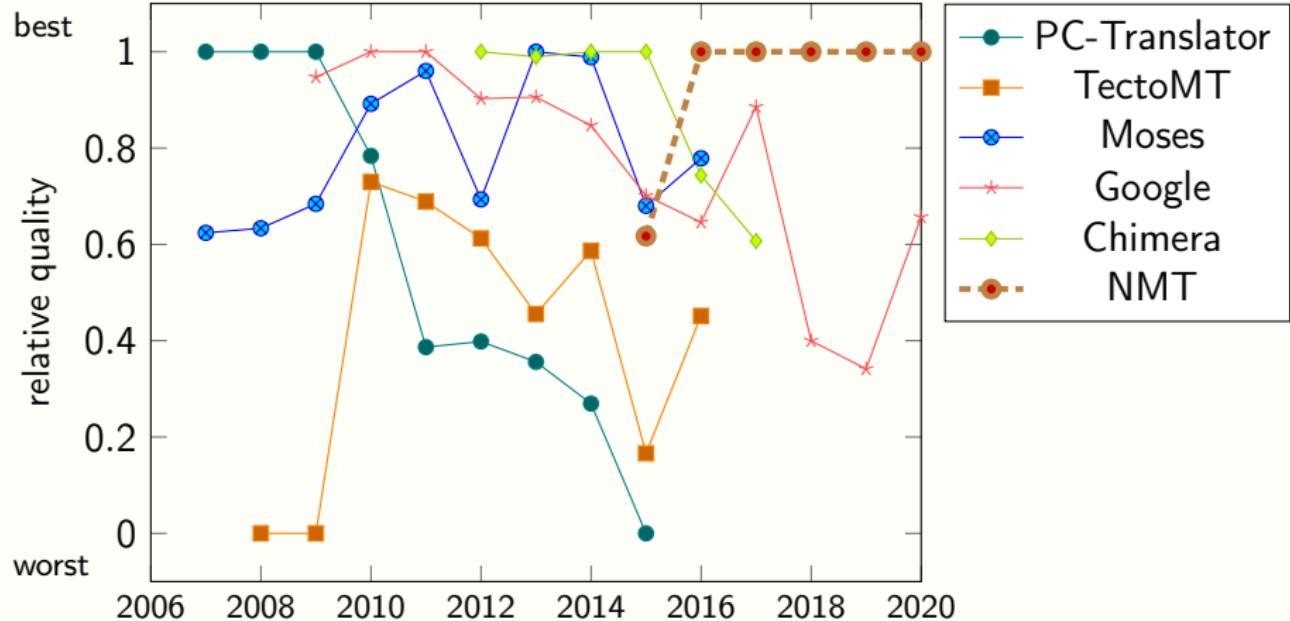
# English→Czech MT in 2007–2020

10



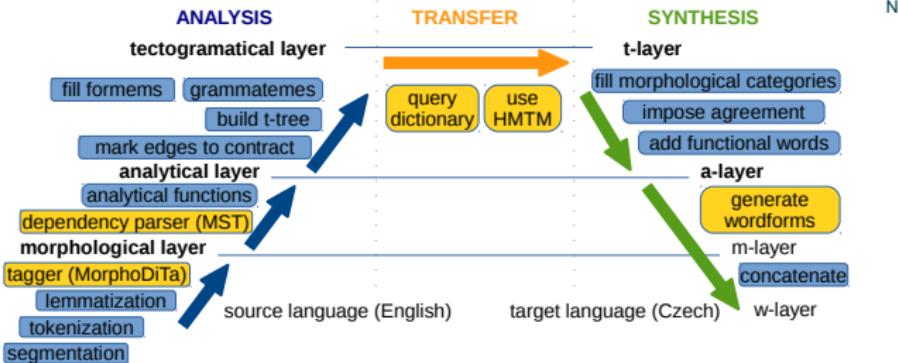
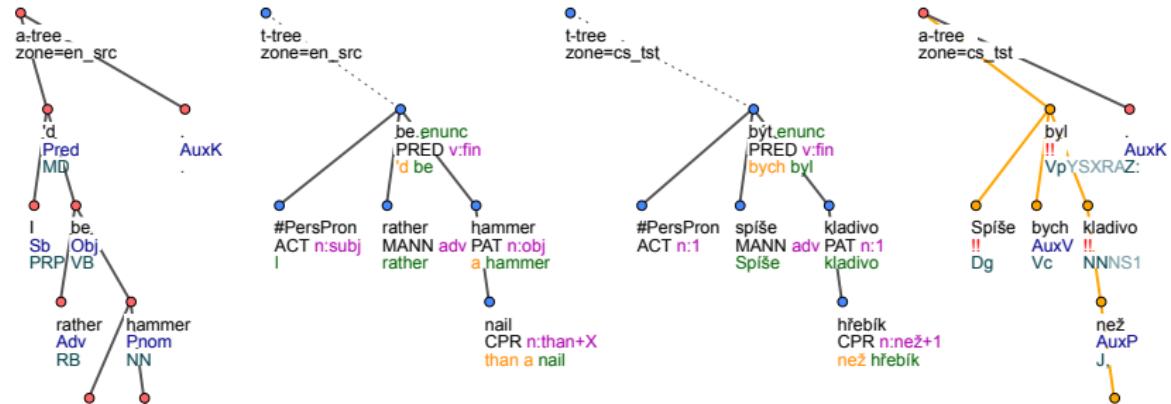
# English→Czech MT in 2007–2020

10



# Deep-syntactic translator TectoMT

11



I'd rather be a hammer than a nail.

Spíše bych byl kladivo než hřebík/nehet.

output_label=hřebík#N	
feature	$\lambda$
child_formeme_n:in+X=1	1.64
is_member_of_coord=1	1.30
child_formeme_v:fin=1	1.04
next_lemma=down	0.84
is_capitalized=1	0.79
<b>+precedes_parent=0</b>	<b>0.75</b>
tense_g=post	0.74
<b>+voice_g=active</b>	<b>0.66</b>
prev_lemma=drive	0.66
parent_capitalized=1	0.62
formeme=n:from+X	0.60
<b>+prev_lemma=hammer</b>	<b>0.59</b>
child_lemma_few=1	0.55
child_lemma_remove=1	0.54
sempos=n.denot	0.50
next_lemma=and	0.50
formeme_g=v:until+fin	0.49
child_lemma_rusty=1	0.47
...	

# Machine learning: features in the translation model

12

output\_label=hřebík#N

feature	$\lambda$
child_formeme_n:in+X=1	1.64
is_member_of_coord=1	1.30
child_formeme_v:fin=1	1.04
next_lemma=down	0.84
is_capitalized=1	0.79
+precedes_parent=0	<b>0.75</b>
tense_g=post	0.74
+voice_g=active	<b>0.66</b>
prev_lemma=drive	0.66
parent_capitalized=1	0.62
formeme=n:from+X	0.60
+prev_lemma=hammer	<b>0.59</b>
child_lemma_few=1	0.55
child_lemma_remove=1	0.54
sempos=n.denot	0.50
next_lemma=and	0.50
formeme_g=v:until+fin	0.49
child_lemma_rusty=1	0.47
...	

output\_label=nehet#N

feature	$\lambda$
child_formeme_n:poss=1	1.32
child_lemma_finger=1	1.07
child_formeme_n:of+X=1	0.98
precedes_parent=1	0.88
prev_lemma=black	0.77
child_lemma_broken=1	0.76
child_formeme_v:attr=1	0.70
formeme=n:at+X	0.67
formeme_g=n:attr	0.67
child_lemma_long=1	0.67
next_lemma=file	0.60
child_lemma_false=1	0.58
prev_lemma=false	0.58
+number=sg	<b>0.56</b>
formeme=n:obj	0.53
formeme=n:by+X	0.52
...	

What is this?

13



40 GPU (GeForce GTX 1080 Ti, 12 billion transistors)

13



# Artificial intelligence and its subfields

14

artificial intelligence

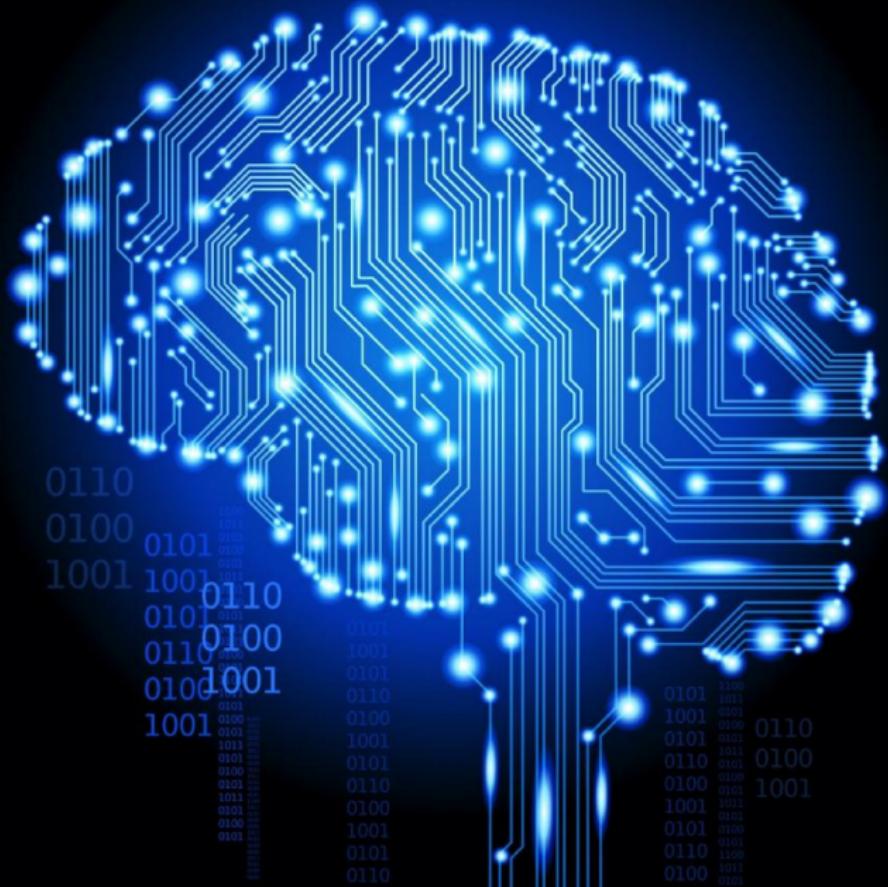
~1950

machine learning

~1980

deep learning

~2010



# Artificial intelligence and its subfields

14

## artificial intelligence

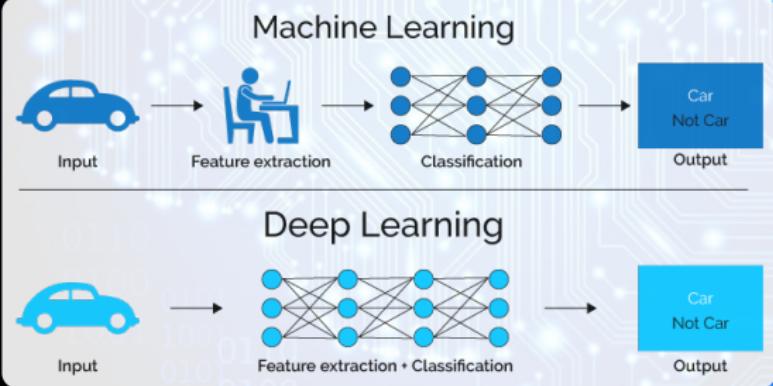
~1950

## machine learning

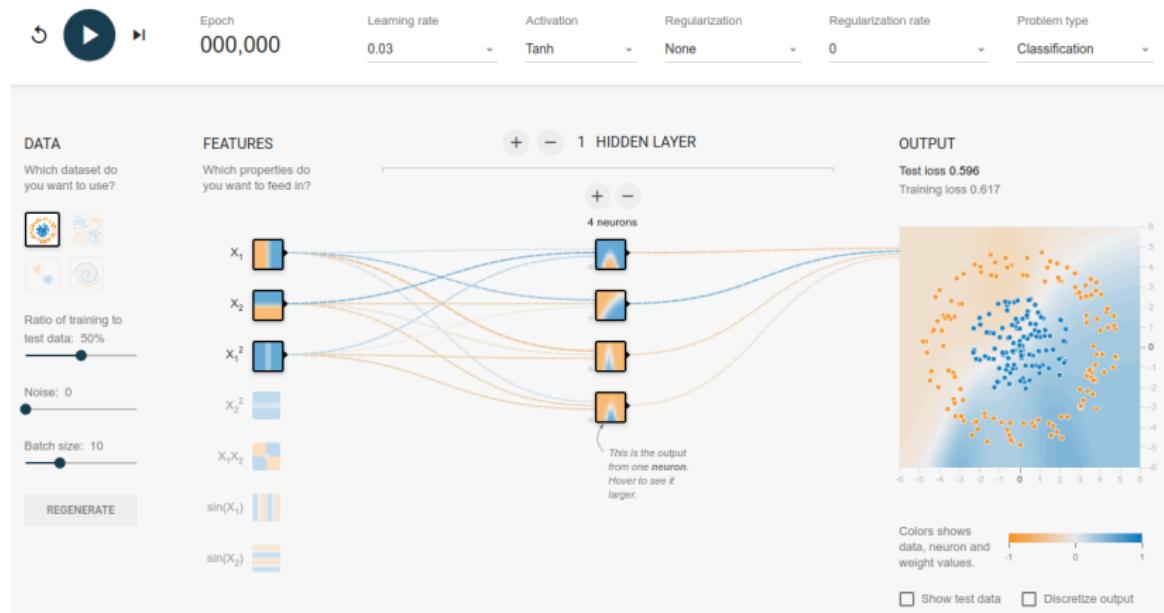
~1980

## deep learning

~2010

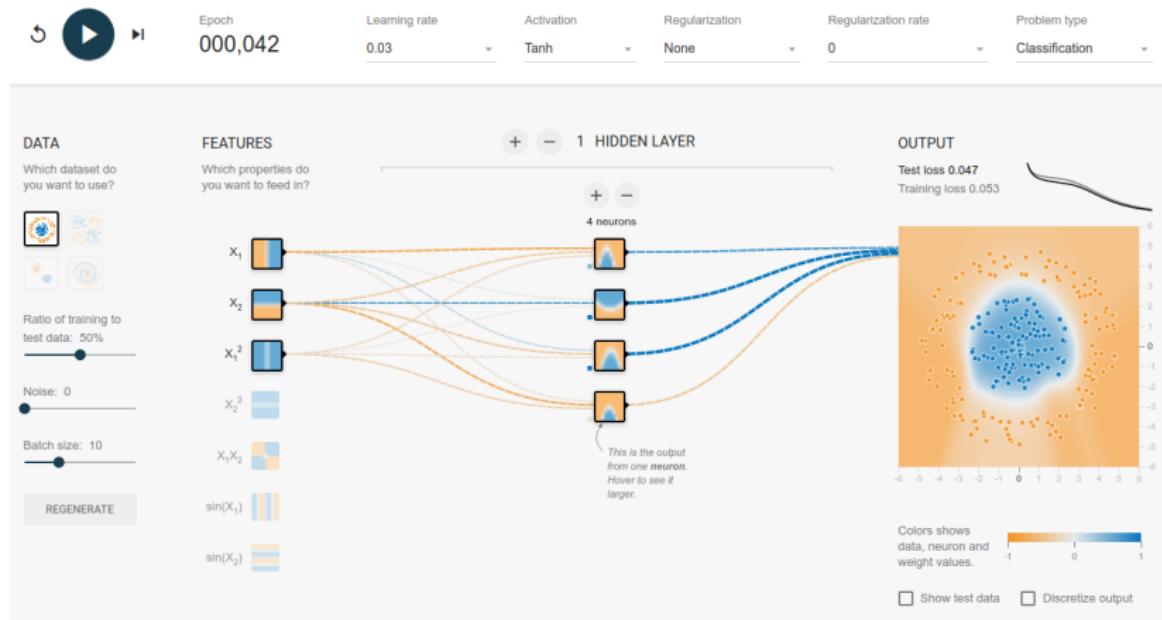


## a simple architecture (16 parameters)



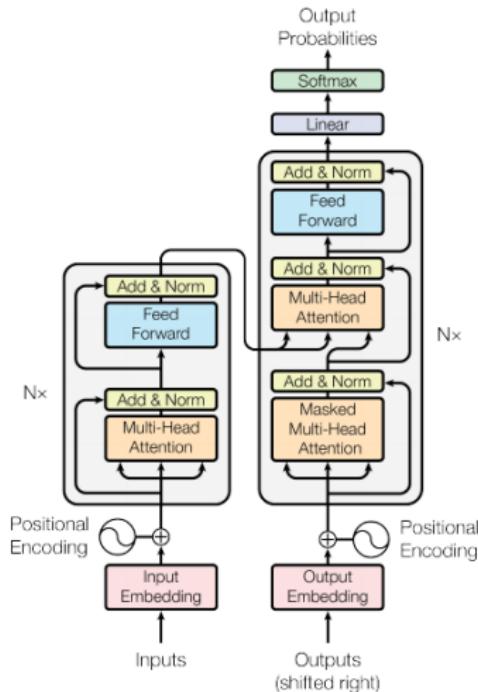
<https://playground.tensorflow.org>

## a simple architecture (16 parameters)

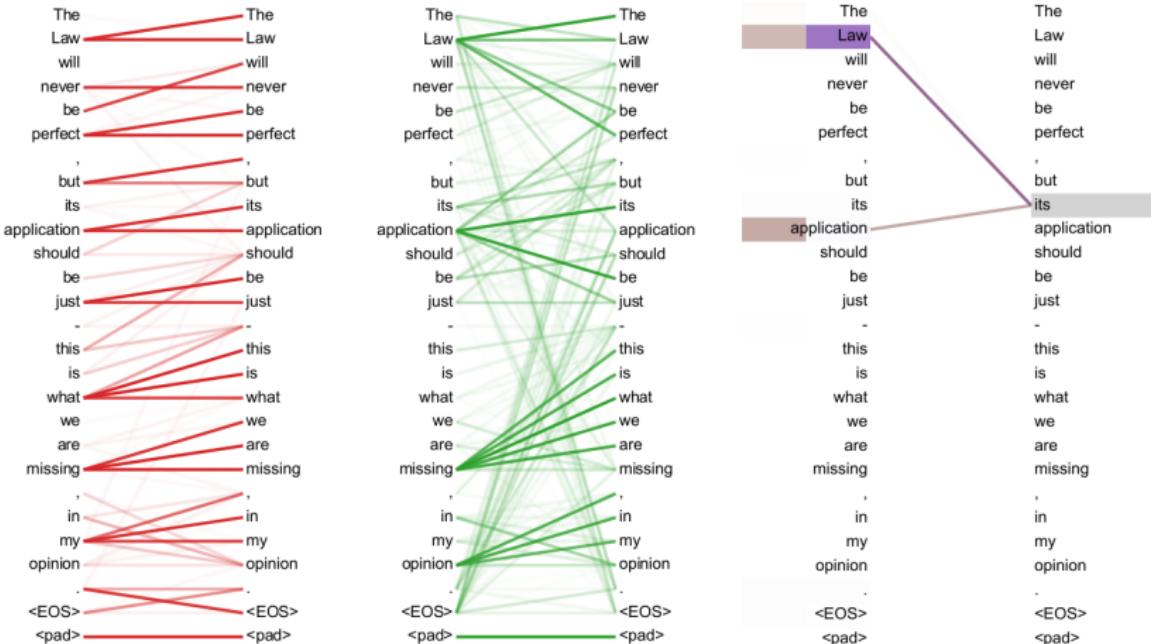


<https://playground.tensorflow.org>

## Transformer architecture (213 millions parameters)

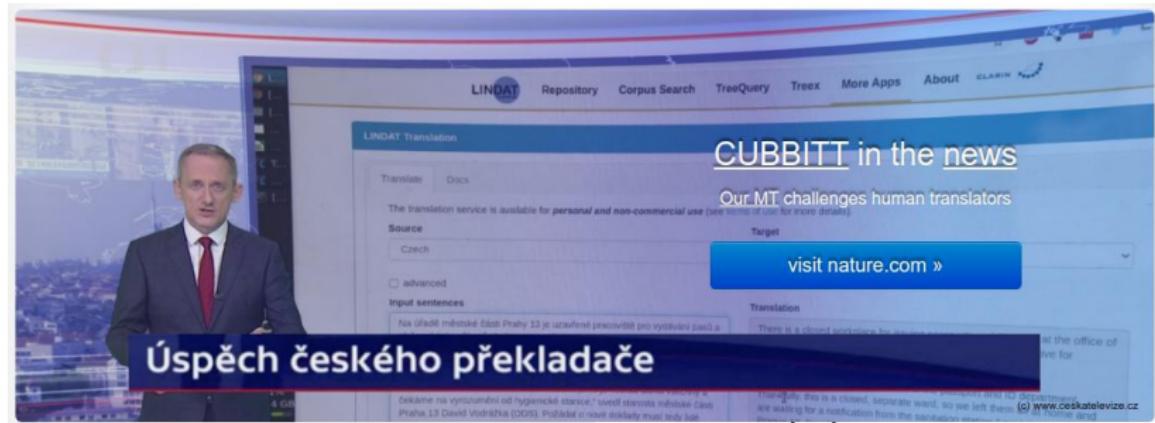


the network learns important relationships between words  
 (in an unsupervised way, via self-attention layers)



# Why CUBBITT made the headlines in 2020?

17



Try CUBBITT at  
<https://lindat.cz/cubbitt>  
(En↔Cs, Fr, Pl)

nature > nature communications > articles > article  
Article | Open Access | Published: 01 September 2020

**Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals**

Martin Popel , Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar & Zdeněk Žabokrtsky

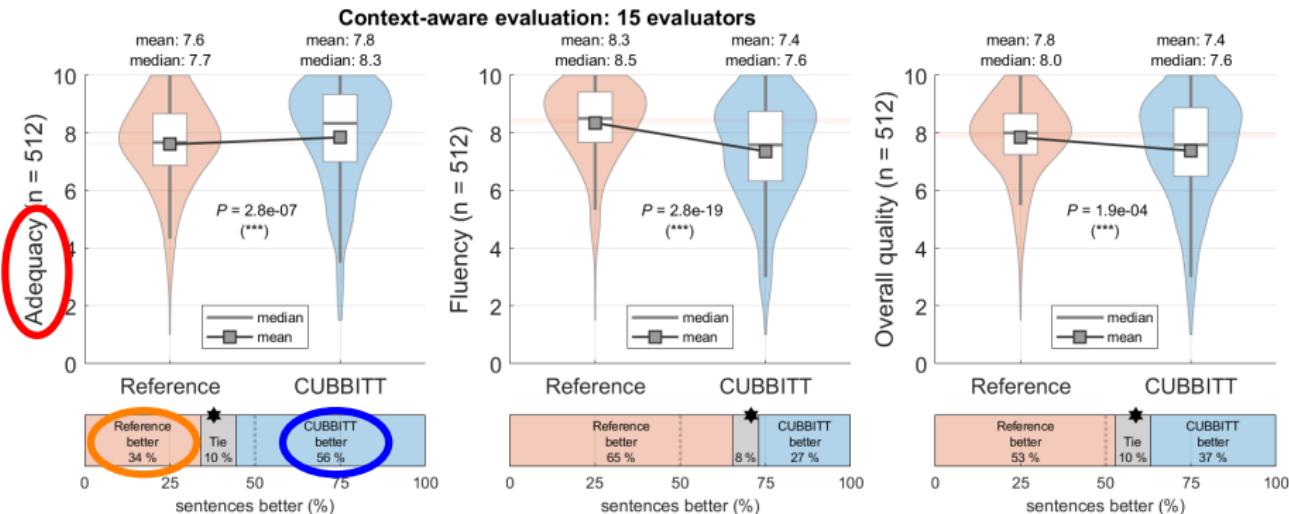
*Nature Communications* 11, Article number: 4381 (2020) | [Cite this article](#)

6273 Accesses | 76 Altmetric | [Metrics](#)

# The main result of our human evaluation

18

56 % sentences translated more adequately by CUBBITT,  
34 % sentences by a professional translation agency



Support The Guardian    Subscribe    Find a job    Sign in

# The Guardian

News   Opinion   Sport   Culture   Lifestyle

World ► Europe US Americas Asia Australia Middle East Africa Inequality

[Facebook](#)

## Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer



zdroj: [The Guardian](#) (2017)

Support The Guardian    Subscribe    Find a job    Sign in

# The Guardian

News   Opinion   Sport   Culture   Lifestyle

World ► Europe US Americas Asia Australia Middle East Africa Inequality

Facebook

## Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer



zdroj: [The Guardian \(2017\)](#)

100+ billion words daily (Google, Microsoft, Baidu, Amazon,...)  
market size in 2020: USD 650 million, but rapidly growing

Source	Jana je žena. Pracuje jako průvodčí.
Google	Jana is a woman. He works as a guide.
Bing	Jana is a woman. He works as a conductor.
CUBBITT	Jane is a woman. He works as a conductor.
CUBBITT-doc	Jana is a woman. <b>She</b> works as a conductor.

Source	Saša je muž. Pracuje jako průvodčí.
DeepL	Sasha is a man. She works as a conductor.
CUBBITT-doc	Sasha is a man. <b>He</b> works as a conductor.

Source	Pavla není muž. Pracuje jako průvodčí.
DeepL	Paul is not a man. He works as a conductor.
CUBBITT-doc	Pavla is not a man. He works as a conductor.

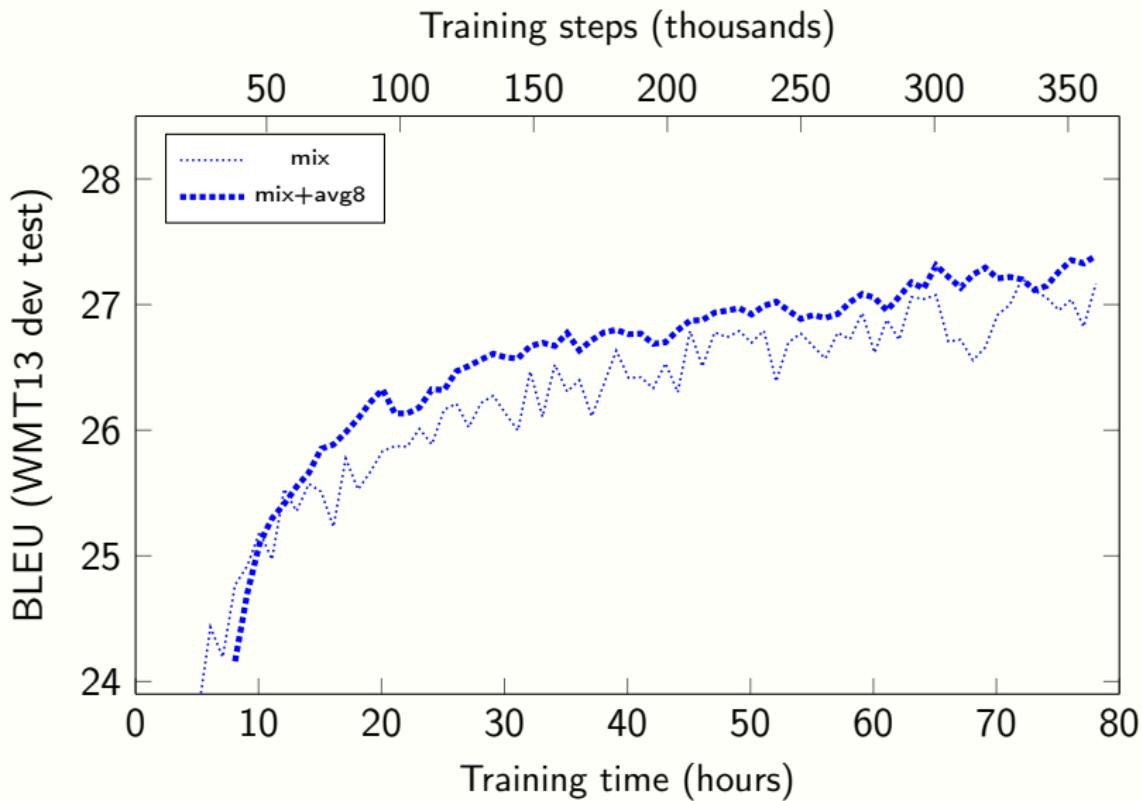
Want to know more?

- For EN→CS translation, we can exploit monolingual CS data.
- Translate the data back to English (with any CS→EN MT).
- Prepare synthetic parallel data (orig-CS, synth-EN).
- Train on both authentic and synthetic
  - **fine-tune BT**: first auth then auth+synth
  - **mix BT**: shuffle auth and synth sentences 1:1



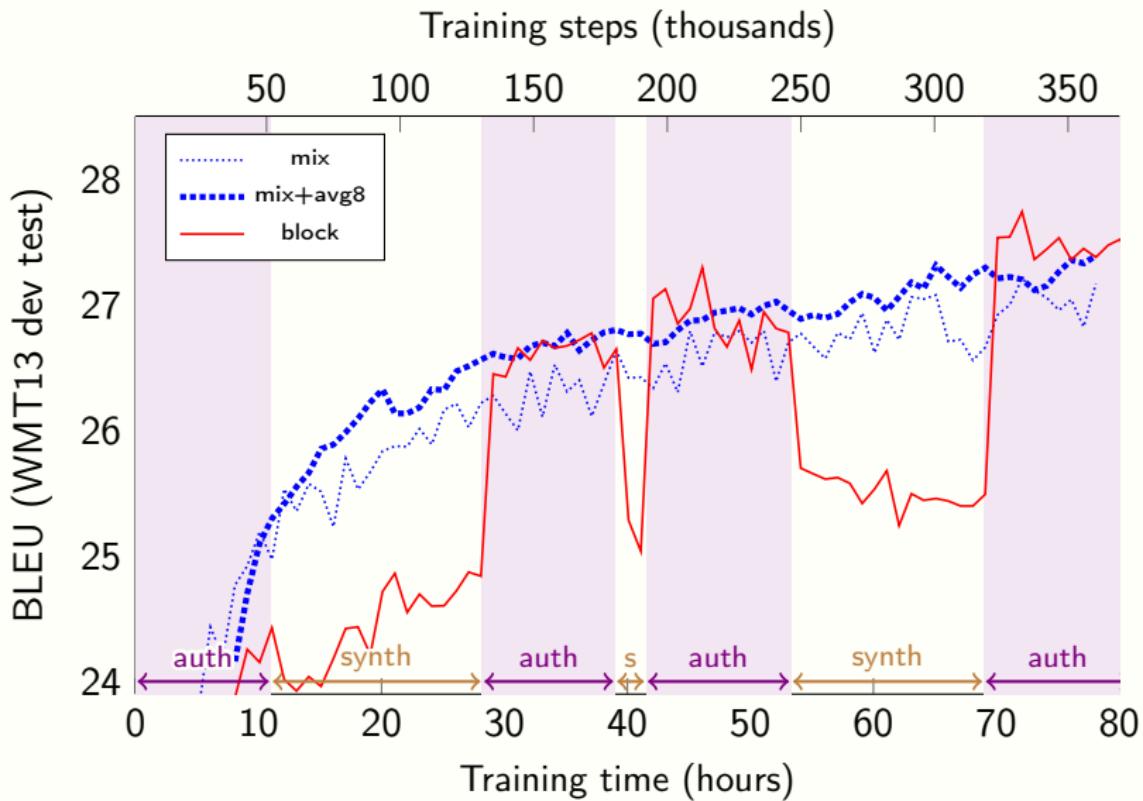
- For EN→CS translation, we can exploit monolingual CS data.
- Translate the data back to English (with any CS→EN MT).
- Prepare synthetic parallel data (orig-CS, synth-EN).
- Train on both authentic and synthetic
  - **fine-tune BT**: first auth then auth+synth
  - **mix BT**: shuffle auth and synth sentences 1:1
  - **block BT**: no shuffle, just concatenate auth and synth blocks





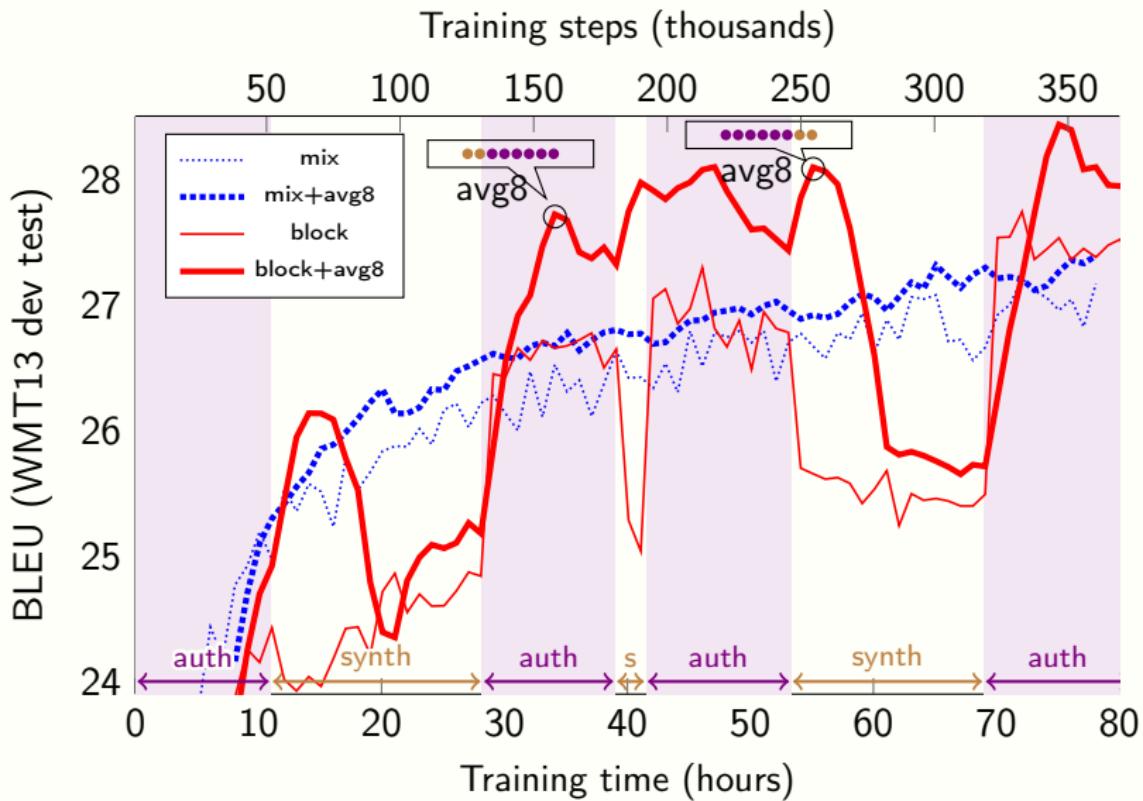
# Block Backtranslation

23



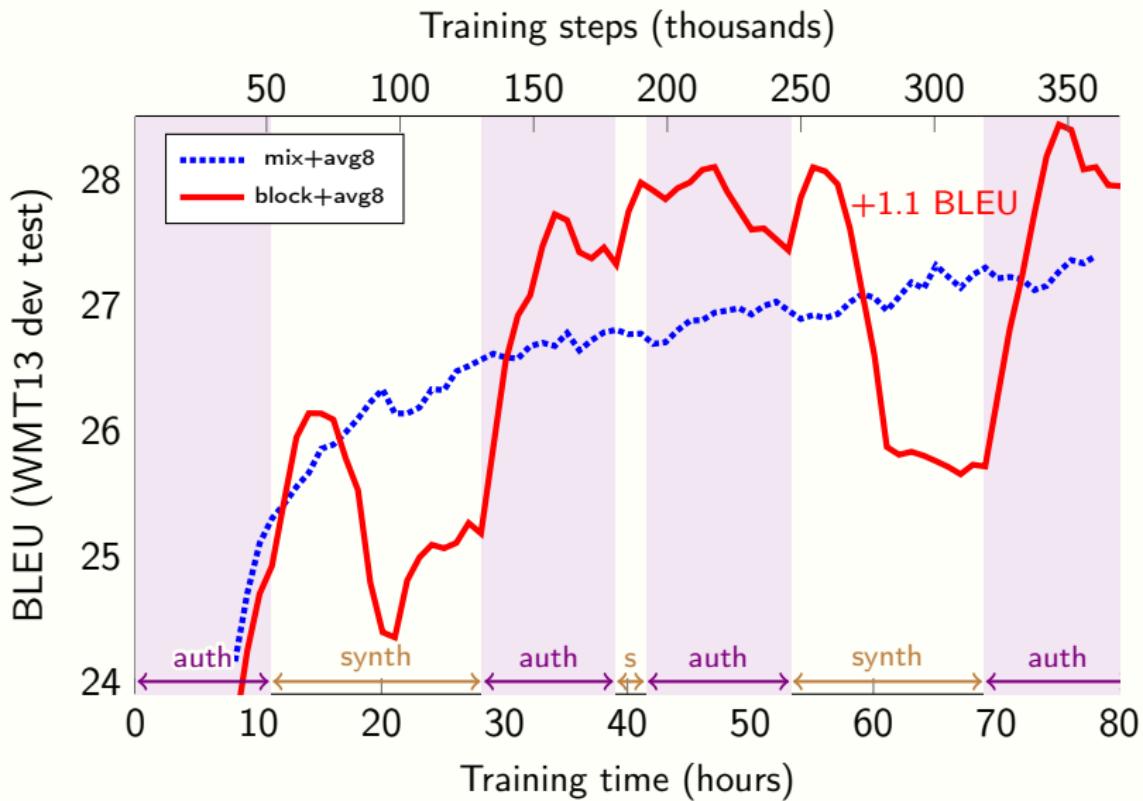
# Block Backtranslation

23



# Block Backtranslation

23



# Manual evaluation example

24

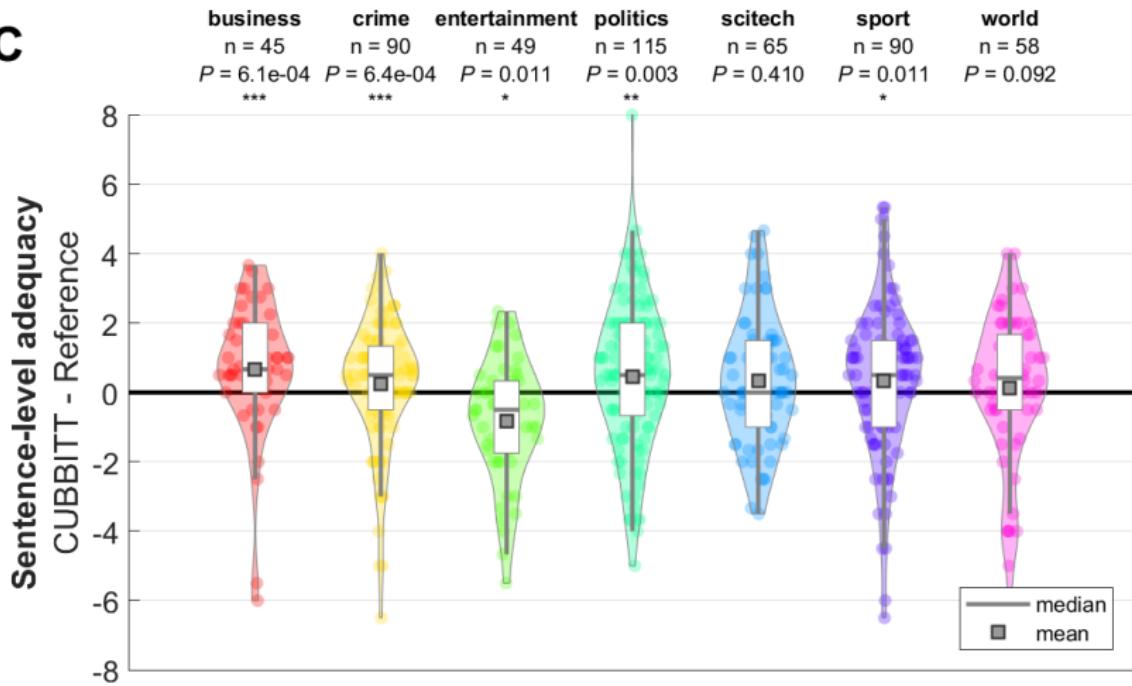
blind, sentence-level but document-aware, side-by-side (**RankME**)

	G	H	I	J	K	L	M	N	O	P
1	Source	Translation1	T1_overall	T1_adeqacy	T1_fluency	Translation2	T2_overall	T2_adeqacy	T2_fluency	Optional comment
168	"And we're protecting our shareholders from employment litigation."									
169	Companies started taking ethics, values and employee engagement more seriously in 2002 after accounting firm Arthur Andersen collapsed because of ethical violations from the Enron scandal, Quinal said.									
170	But it wasn't until "social media came into its own" that companies realized they couldn't stop their dirty laundry from going viral online.									
171	"Prior to using technology to monitor ethics, people used hope as a strategy," he said.									
172	Both Glint and Convergent offer their software as a service, charging companies recurring fees to use their products.									
173	It's a business model and opportunity that has the approval of venture capital investors, who have propped up both start-ups.	Je to obchodní model a příležitost, kterou schvaluji odvážní kapitáloví investoři, jenž podporují oba startupy.	7	6	7	Je to obchodní model a příležitost, která má souhlas investorů rizikového kapitálu, kteří podporují oba start-upy.	10	10	10	T1: chybný překlad termínu "venture capital"
174	Convergent raised \$10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to \$47 million.	Convergent vybral v rámci své únorové kampaně od frenů jako Sapphire Ventures a Tola Capital celkově 10 milionů \$. A nakonec si odnesl kapitál ve výši 47 milionů \$.	3	4	3	Convergent získal v únoru finanční prostředky ve výši 10 milionů dolarů od firem jako Sapphire Ventures a Tola Capital, čímž se jeho celkový kapitál zvýšil na 47 milionů dolarů.	10	10	10	
175	Glint secured \$10 million in November from Bessemer Venture Partners, bringing its total funding to \$60 million.	Glint získal v listopadu 10 milionů \$ od Bessemer Venture Partners a v průběhu celé kampaně získal 60 milionů \$.	5	4	5	Glint získal v listopadu 10 milionů dolarů od společnosti Bessemer Venture Partners, čímž jeho celkové financování dosáhlo 60 milionů dolarů.	10	10	10	
176	These investments hardly come as a surprise, given the interconnected nature of companies, culture and venture capital.	Tyto investice jsou stěží překvapující vzhledem k vzájemné povaze společnosti, kultury a rizikovému kapitálu.	3	4	3	Tyto investice nejsou vzhledem k propojenosti společnosti, kultury a rizikového kapitálu zádným překvapením.	10	10	10	
177	There's a growing body of research showing today's employees expect more from their workplaces than before.	Narůstající počet výzkumů jasně potvrzuje, že dnešní zaměstnanci očekávají od svého pracoviště více než kdy dříve.	5	5	5	Roste množství výzkumů, které ukazují, že dnešní zaměstnanci očekávají od svých pracovišť více než dříve.	10	10	10	
178	In competitive markets such as Silicon Valley, high salaries and interesting projects are merely table stakes.	A na konkurenčních trzích, jakým je např. Silicon Valley, jsou hlavní výhodou vysoké platy a zajímavé projekty.	6	5	8	Na konkurenčních trzích, jako je Silicon Valley, jsou vysoké platy a zajímavé projekty pouhými sázkami u stolu.	7	8	7	problém: význam terminu "table stakes"
179	Employees want to feel that they're accepted and valued and that they're giving their time to a company with a positive mission.	Zaměstnanci chtějí vnímat, že jsou přijímáni a oceňováni a že věnují svůj čas společnosti, která usiluje o pozitivní poslání.	9	9	9	Zaměstnanci chtějí mít pocit, že jsou přijímáni a ceněni a že věnují svůj čas společnosti s pozitivním posláním.	10	10	9	

# It depends on the input text domain

25

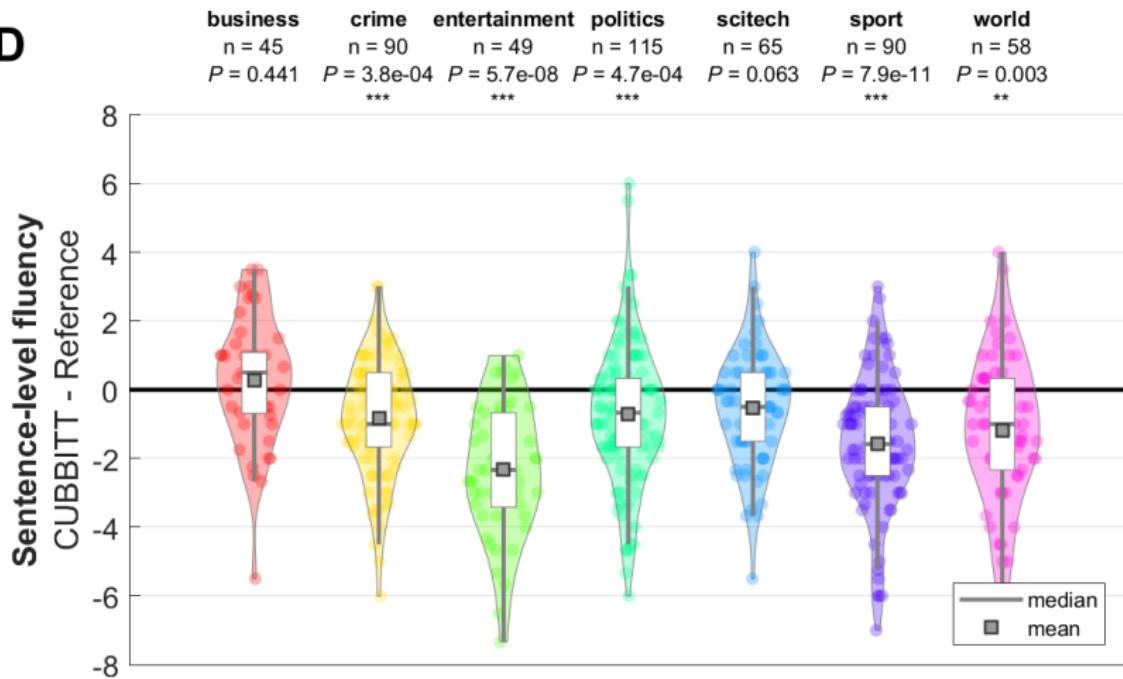
C



# It depends on the input text domain

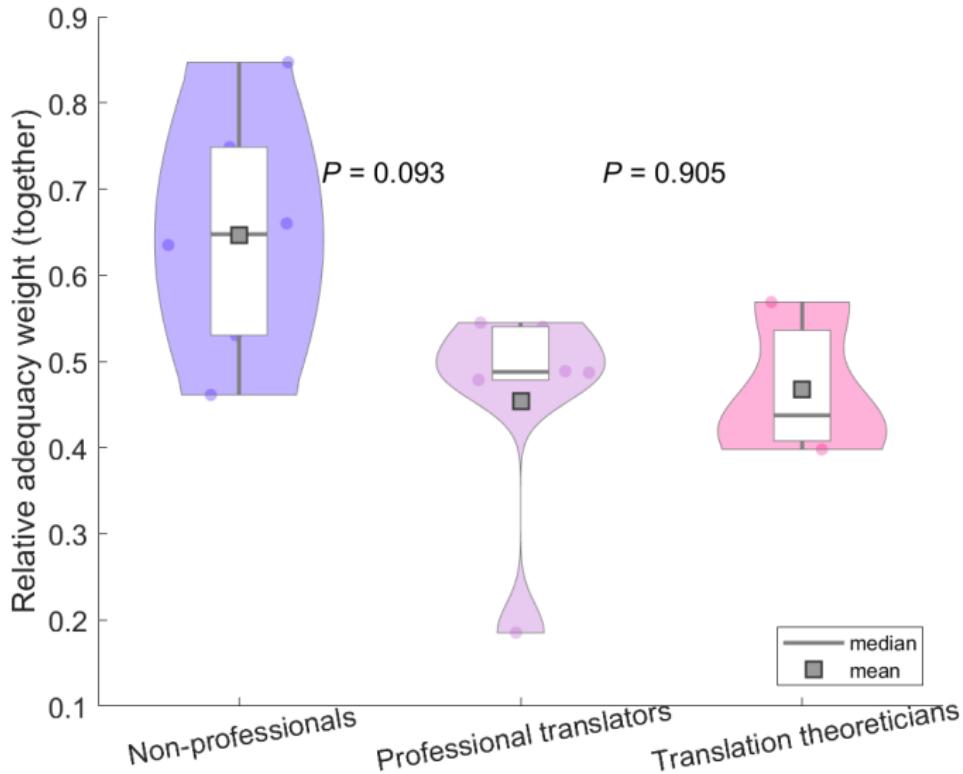
25

D

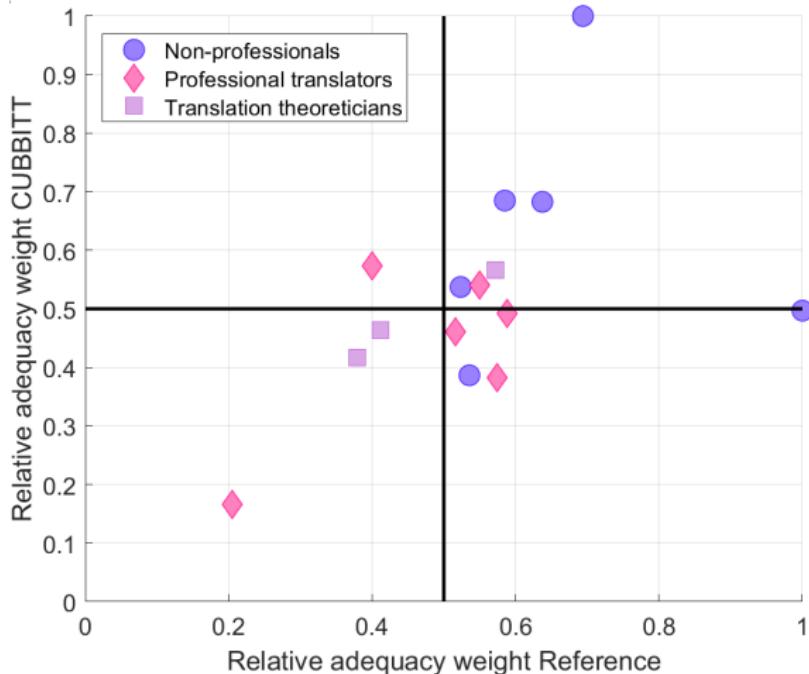


Can we predict the **overall quality**  
as a weighted average of **adequacy** and **fluency**?

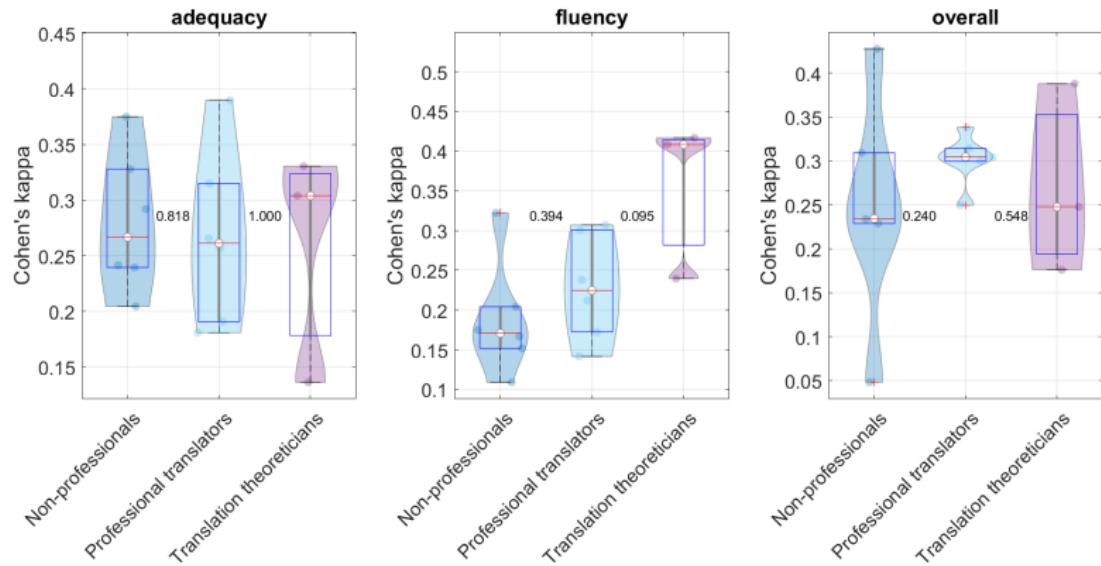
Translators put more emphasis on fluency,  
non-professionals on adequacy.



Some annotators put all the emphasis on adequacy,  
but only for one of the systems.

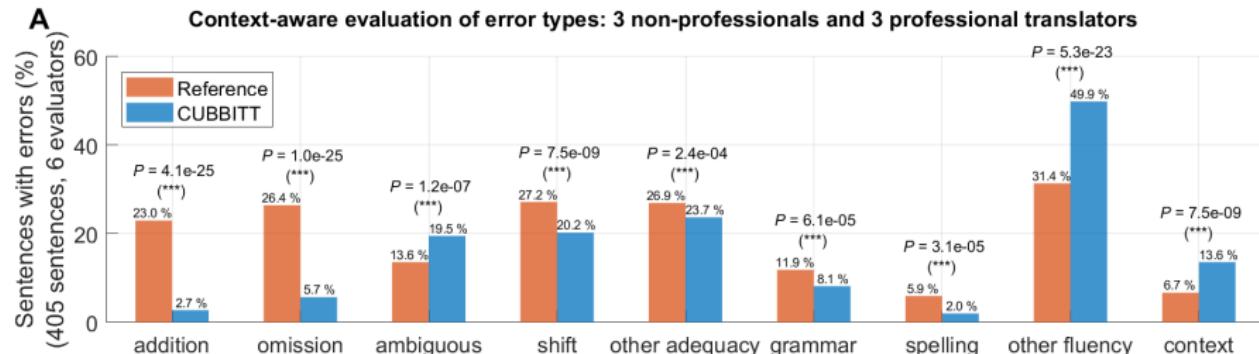


We cannot conclude that professionals are more reliable.



CUBBITT makes more errors than humans in

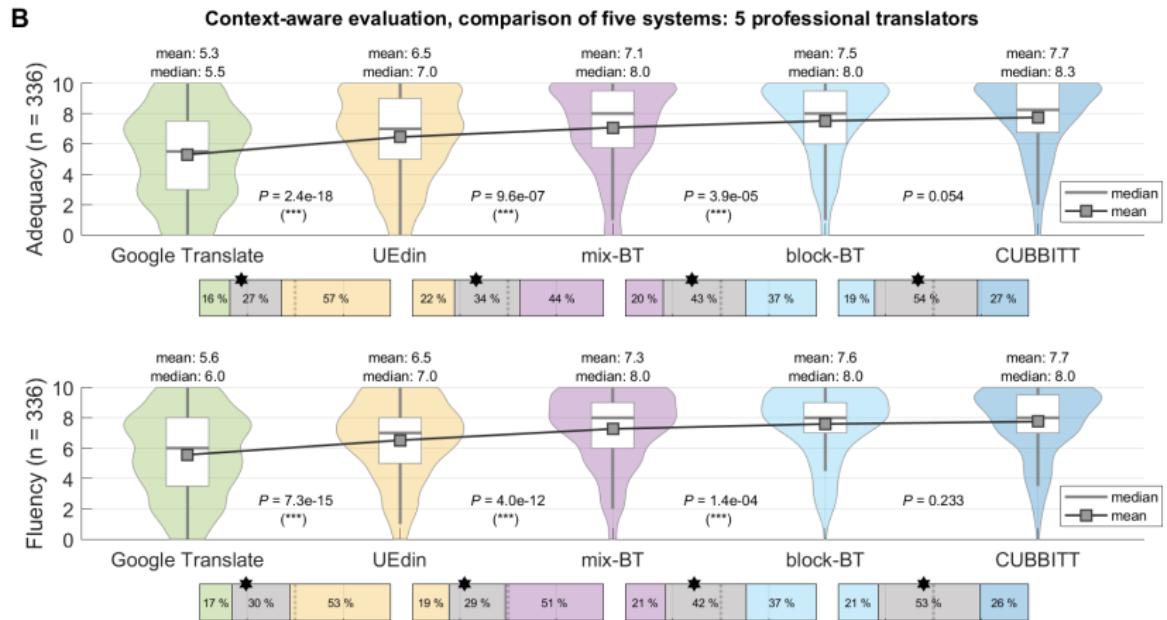
- translation of ambiguous words
- fluency (but not spelling and grammar) and
- cross-sentence context



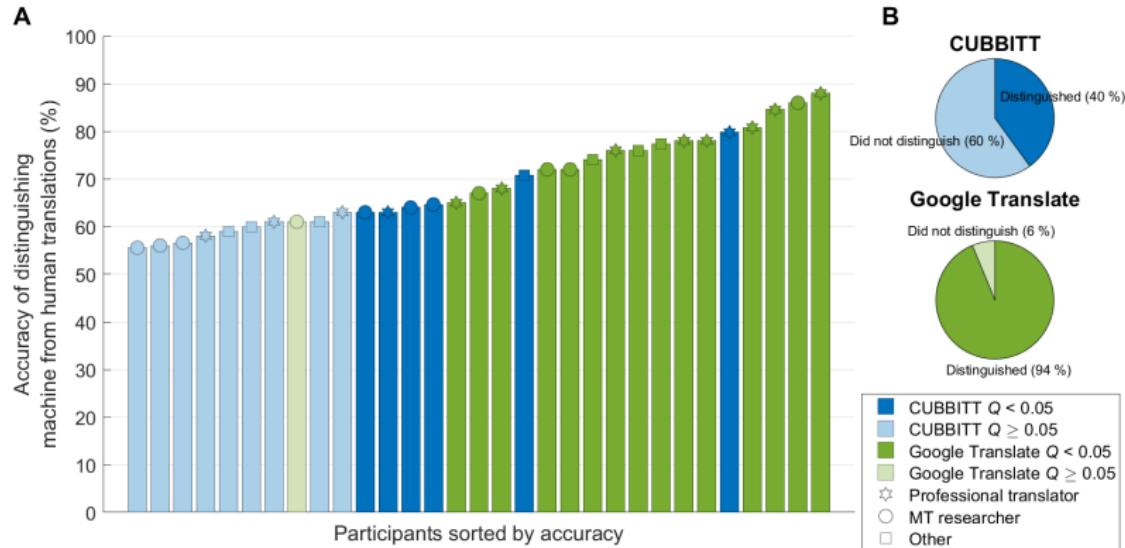
# Ablation analysis

31

BlockBT improves the quality (+0.4 adequacy, +0.3 fluency).



60 % of participants did not distinguish CUBBITT from human translations (on 100 isolated sentences)

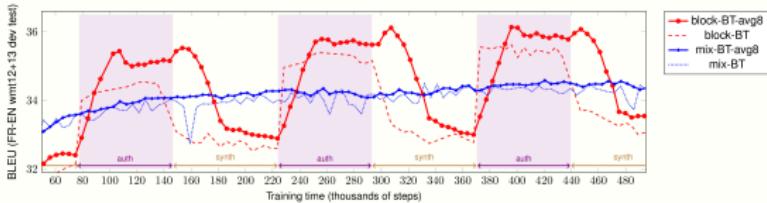


# BLEU Training curves for en-fr and en-pl

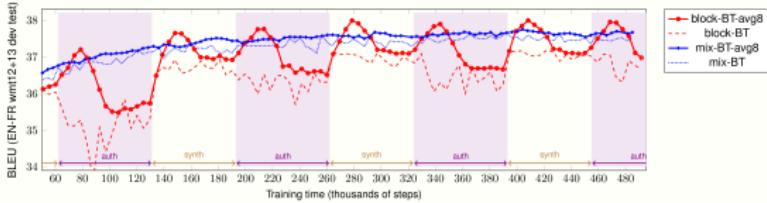
33

**A**

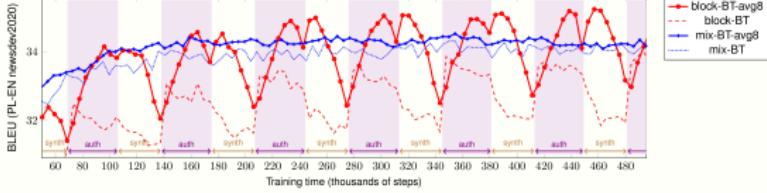
French → English

**B**

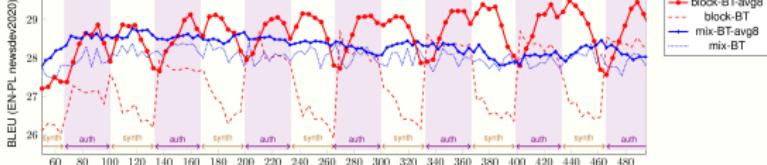
English → French

**C**

Polish → English

**D**

English → Polish



# Why Block-BT works?

34

**A**

Source sentence: "He was an original **guy** and lived life to the full" said **Gray** in a statement.



Translation of block-BT-Avg: "Byl to originální **chlap** a žil život naplno," uvedl **Gray** v prohlášení.

**Checkpoint 1 (SYNTH):** "Byl to originální **chlap** a žil život naplno" uvedl **Šedivý** v prohlášení.

**Checkpoint 2 (SYNTH):** "Byl to originální **chlap** a žil život naplno" uvedl **Šedivý** v prohlášení.

**Checkpoint 3 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

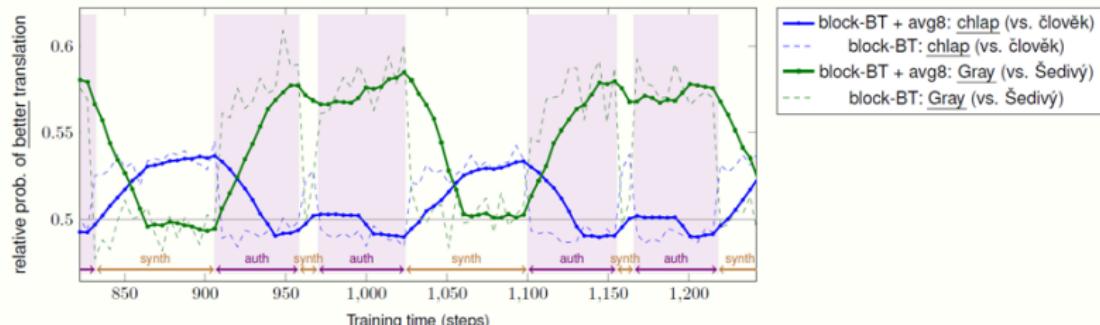
**Checkpoint 4 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 5 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 6 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 7 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**Checkpoint 8 (AUTH):** „Byl to originální **člověk** a žil život naplno,“ uvedl **Gray** v prohlášení.

**B**

source	As good be an addled egg as an idle bird.
Bing	Jako dobrý být popletený vejce jako nečinný pták.
Google	Jako dobrá být včleněná vejce.
T2009	Dobré je fetácké vejce jako činný pták.
T2018	Dobří bud'te plete vejce jako nečinný pták.
CUBBITT	Stejně dobré je být pomateným vejcem jako zahálejícím ptákem.

source	A miss by an inch is a miss by a mile.	Birds of a feather flock together.
Bing	Miss o palec je Miss o míli.	Ptáci peří stáda dohromady.
Yandex	Slečna tím, že palec je vedle o míli.	Vrána k vráně sedá.
Google	Chybějící palcem je míle vzdálená míle.	Vrána k vráně sedá.
T2009	Slečna palec je slečna milionu.	Ptáci v bederním hejnu spolu.
T2018	Slečna palce je slečna míle.	Ptáci péřového hejna spolu.
CUBBITT	Minutí o centimetr je o kilometr.	Vrána k vráně sedá.

Try CUBBITT at: <https://lindat.cz/cubbitt>