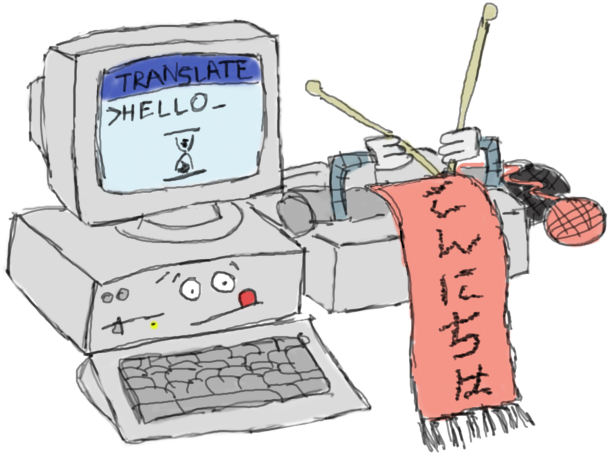
Machine translation and artificial intelligence

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 dříči.

Natural Language Processing?

Spell check vs. grammar check

Chlapci šly.
Chlapec šli do školy.

Spell check vs. grammar check

Dívce nešly hodinky.

Chlapci šly.

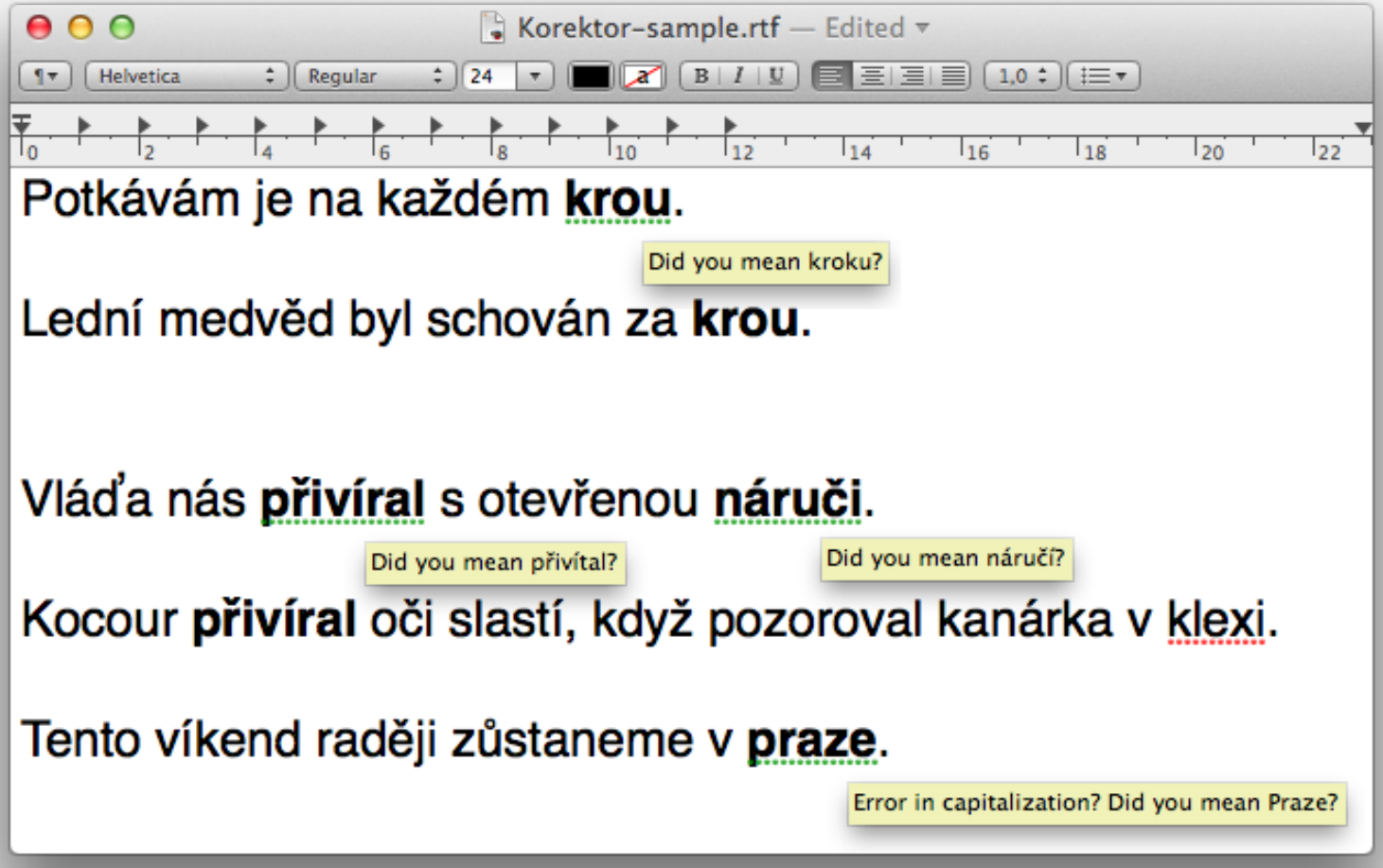
Chlapec šli do školy.

Dívce nešly hodinky. Chlapci šly.

Kdo kam co donesl? Chlapec šli do školy.

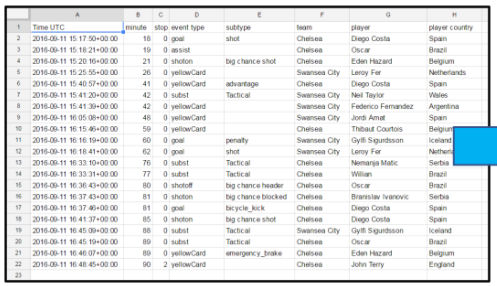
Spell check vs. grammar check

http://ufal.cz/korektor

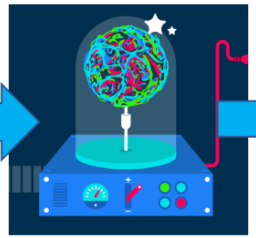


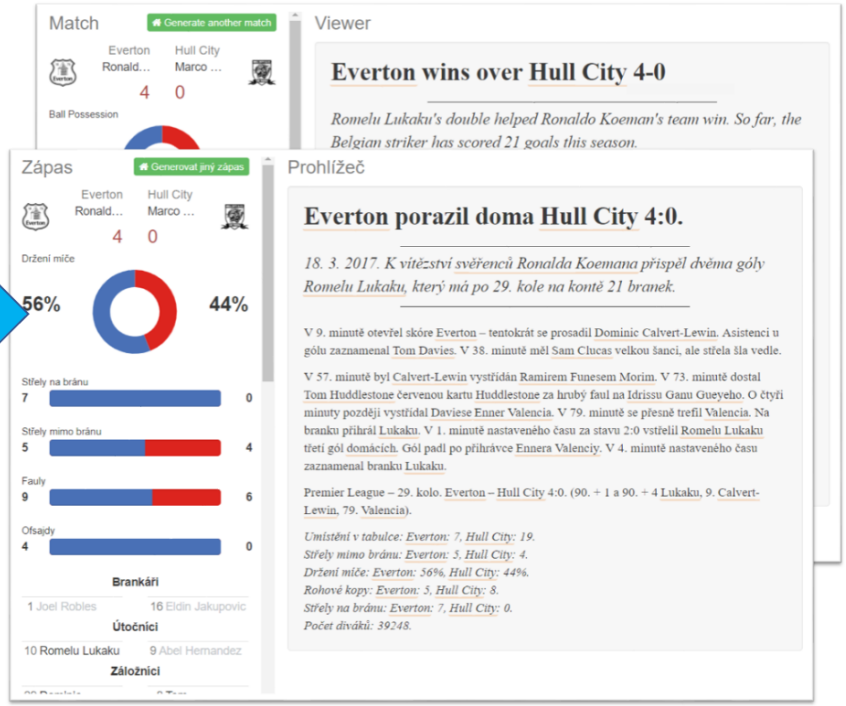
Automatic news generation

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



Structured data
(real-time & historical)





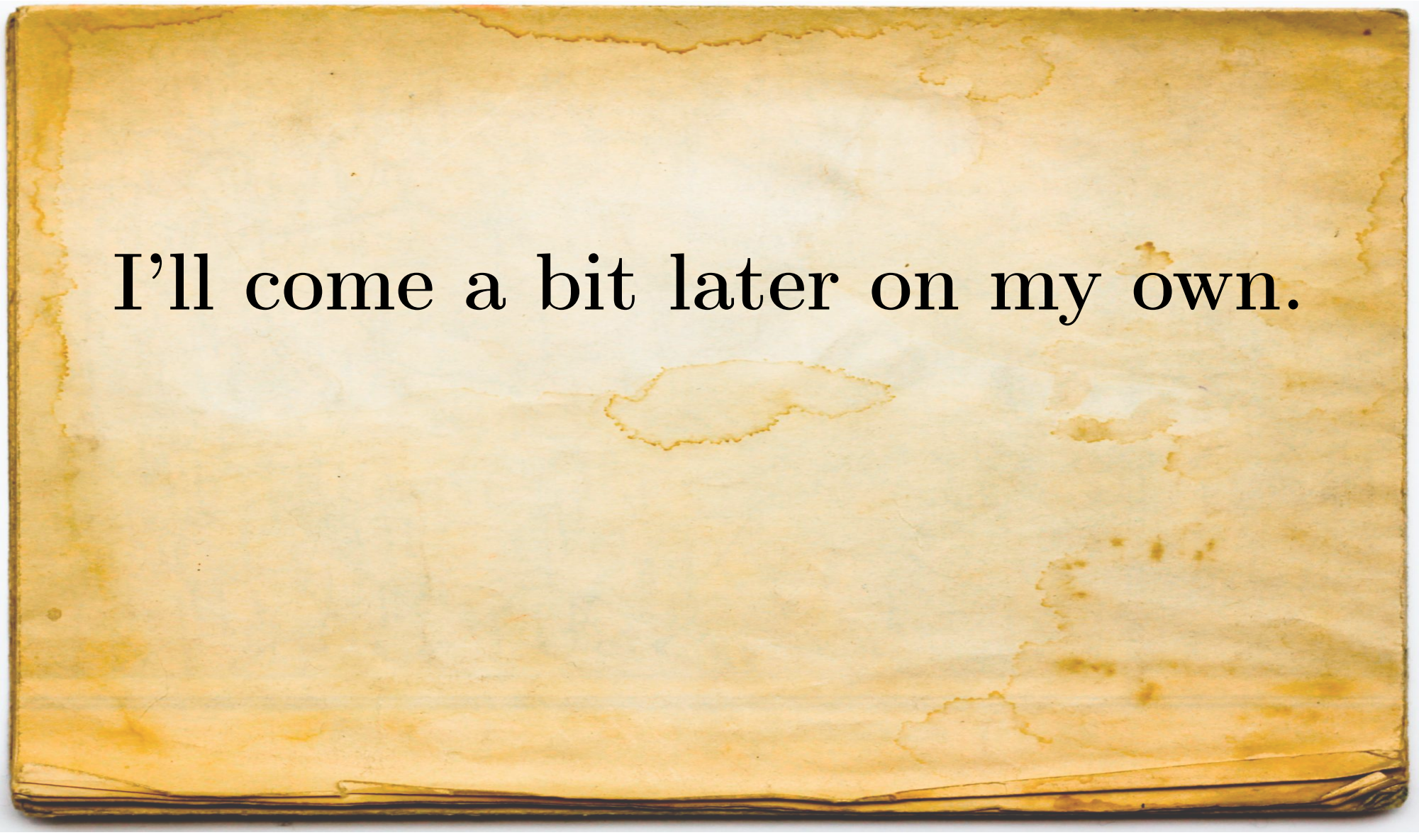
Article / Report

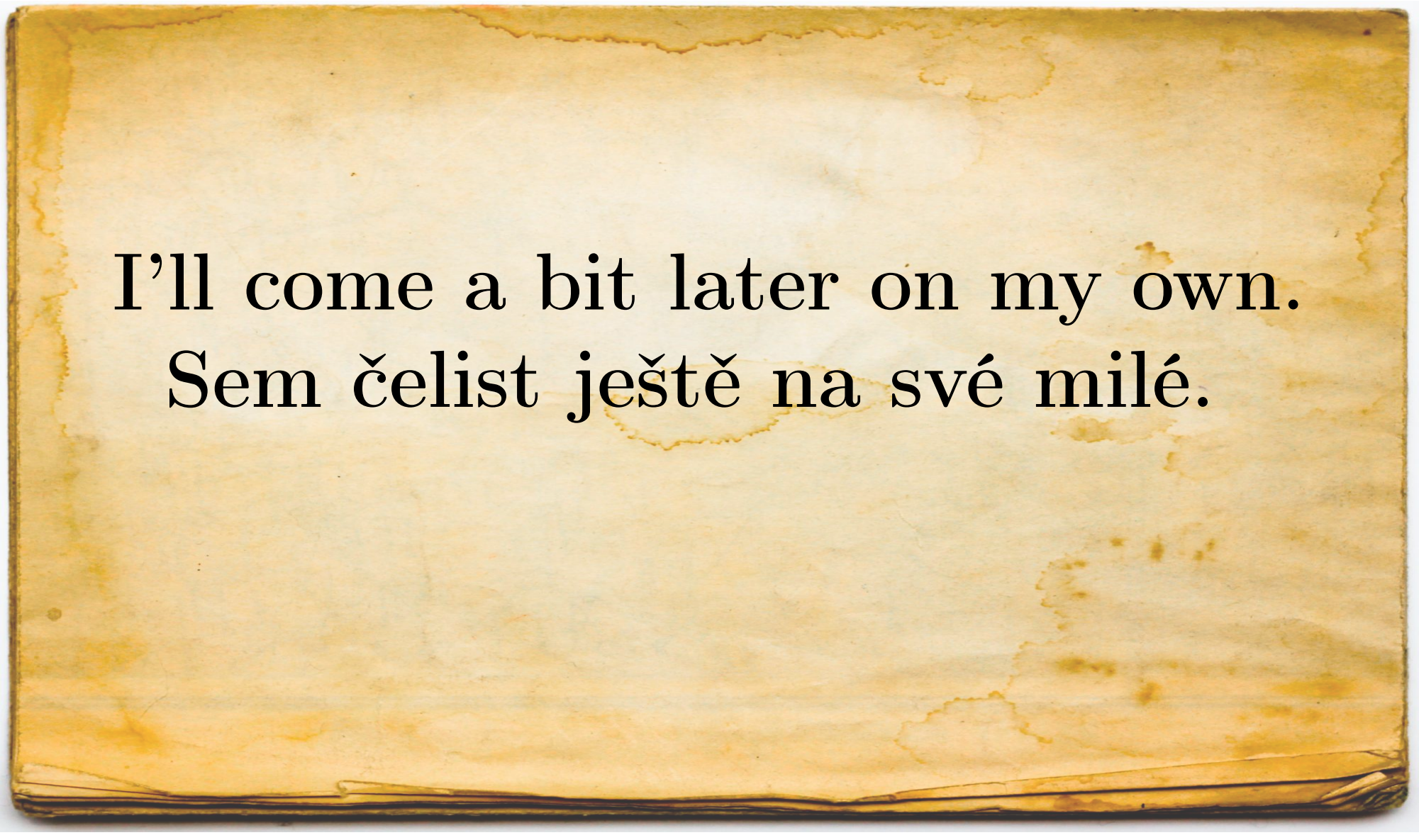
Byl by to rytíř, kde v pláně hřích vzlet,
Vědě jsem jse seheldo na přídoutně v světě si nezastavá:
„Ukryjemné, chvěla, milý nás jest

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.

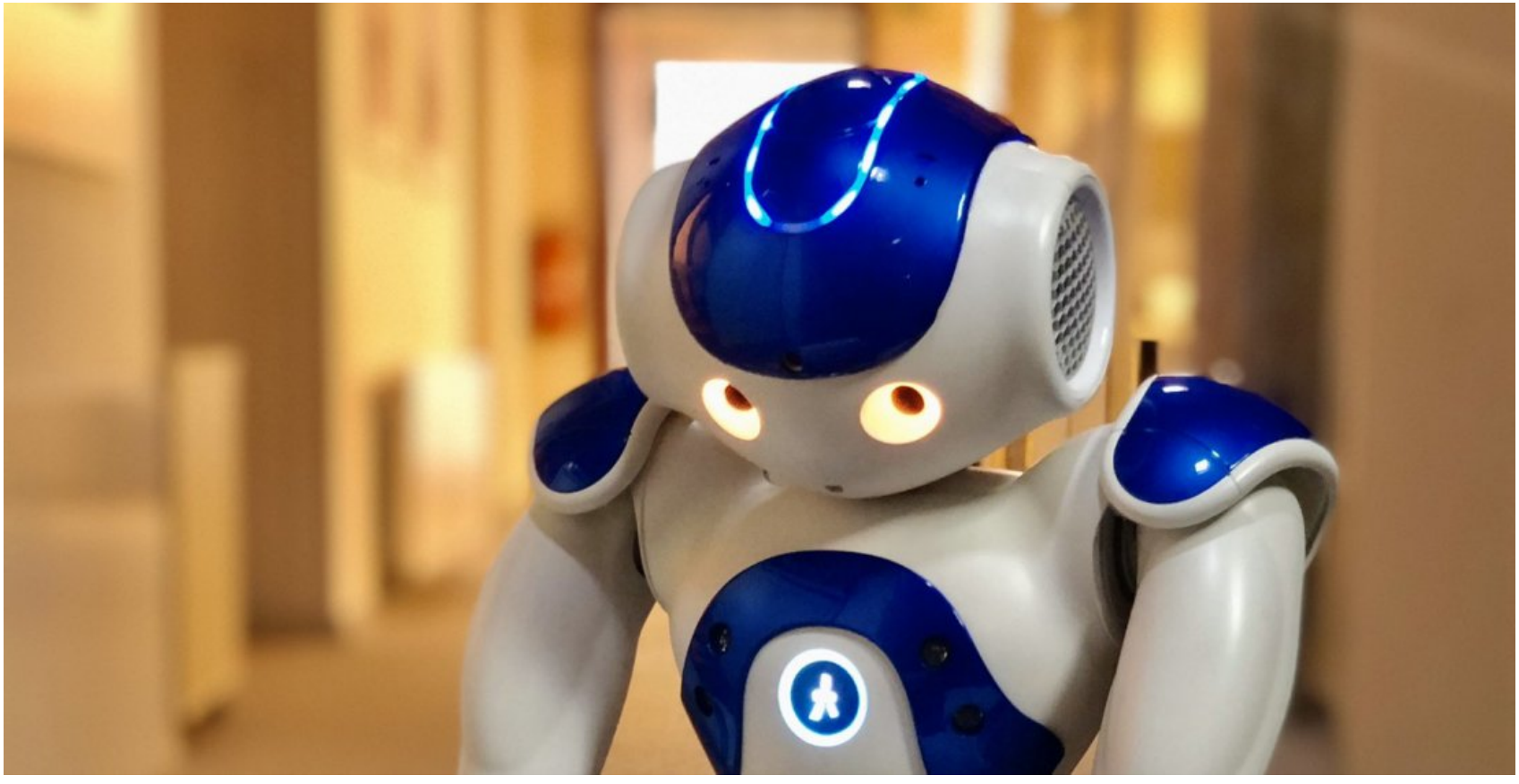

I'll come a bit later on my own.


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

Can a robot write a theatre play?

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. Prolomí kódovanou řeč i šifrovací aplikace

KATEŘINA KOLÁŘOVÁ

Multidisciplinární tým. Vyšetřování podezření z korupce nebo projevů sexuálního harašení na pracovišti spojuje práci rozdílných vědních oborů. Analytický tým vede absolventka matematicko-fyzikální fakulty Kateřina Veselovská (vlevo), právníky advokátka Jaroslava Kračúnová (vpravo). FOTO MAFRA – DAN MATERNA

Ve druhém patře moderní pražské kancelářské budovy Nile House usedá k jednacímu stolu pestře složený tým odborníků. Právníci, forenzní analytici a vyšetřovatelé zjišťují, jestli byly v prošetřované společnosti globálního významu uzavřeny pro firmu nevýhodné smlouvy. Experti společnosti Deloitte zkoumají, jestli zaměstnanci vyšetřované společnosti „šli na ruku“ dodavateli služeb, a za úplatek mu umožnili vyhrát lukrativní zakázku. K vyšetřování používají specialisté techniku automatické analýzy dat, která prohlédne i kódovanou řeč.

„Velmi často řešíme právě vztahy dodavatelů s nákupním oddělením, ty jsou problematické téměř v každé prošetřované společnosti,“ vysvětluje Jaroslava Kračúnová, advokátka a partnerka kanceláře Ambuz & Dark Deloitte Legal, jež spolupracuje s forenzním týmem vyšetřovatelů. Multidisciplinární tým už vyšetřoval i podezření z financování terorismu v zahraničí. „Takto závažné trestné činy se velmi těžko prokazují izolovaně u prošetřované společnosti klienta. Proto často, paralelně s naším šetřením, probíhá i policejní vyšetřování,“ říká Kračúnová.

Rychlý zásah

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

Rozpoznání kódované řeči

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

„Na základě blízkosti slovní záso-

> 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 textový soubor.

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

Kateřina Veselovská

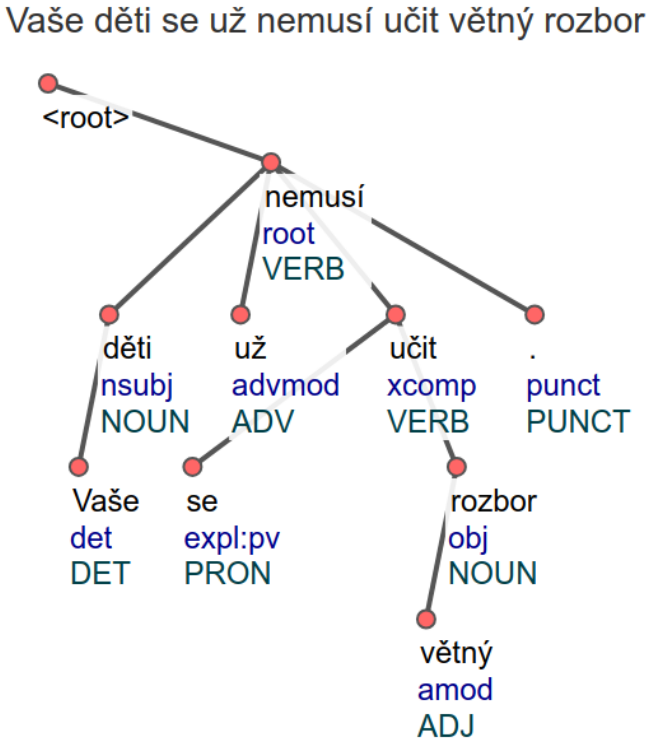
- Manažerka oddělení Data Analytics Deloitte, kde vede tým pro analýzu nestrukturovaných dat. Absolvovala doktorandský program na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze. Věnovala se vývoji softwaru pro textovou analytiku a business poradenství v projektech týkajících se oblasti nestrukturovaných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

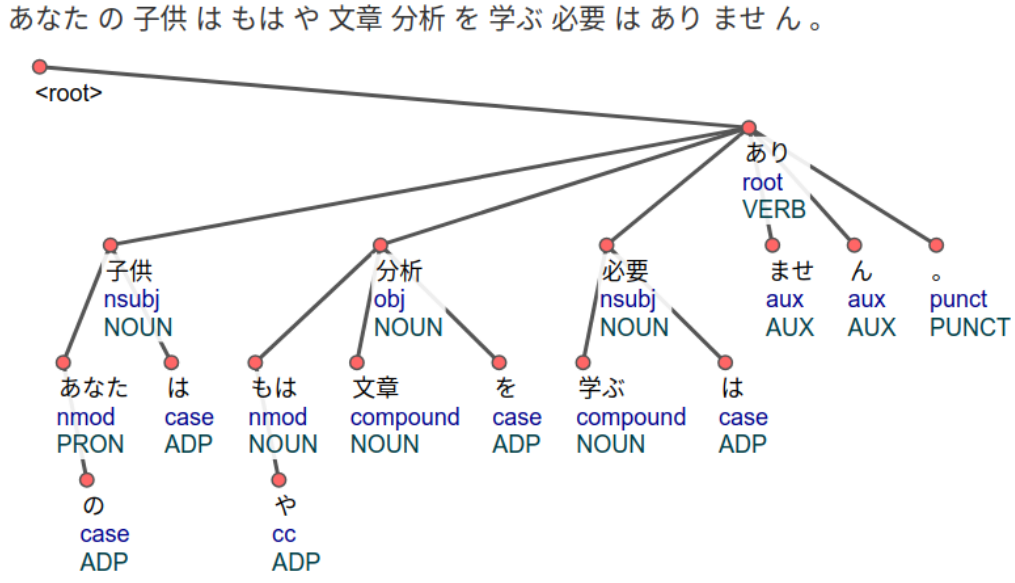
Jaroslava Kračúnová

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

Automatic syntactic analysis

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





https://lindat.cz/services/udpipe/

Can you add and subtract numbers?

What about words and images?

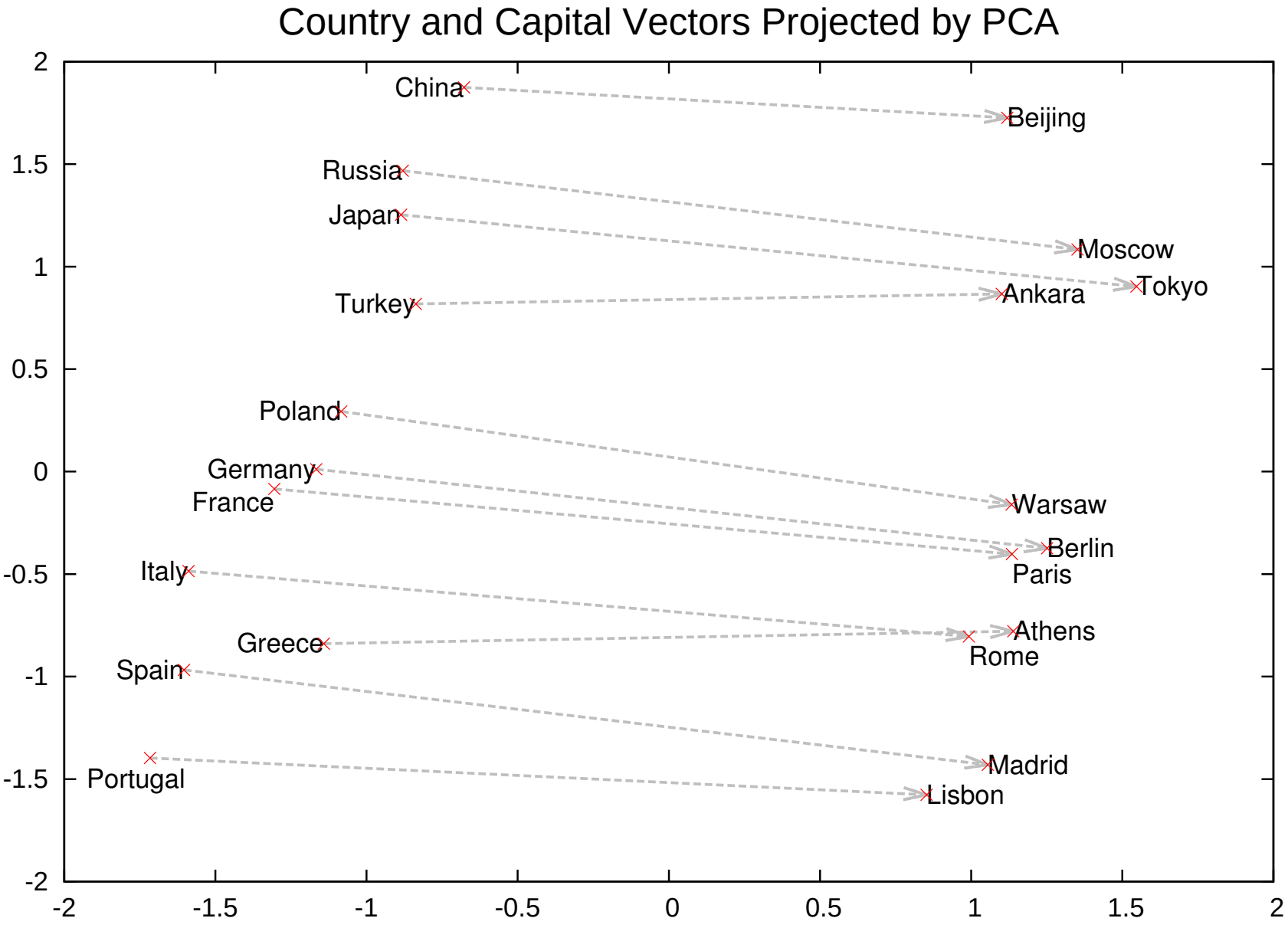
king - man + woman = ?

king - man + woman = queen

Tomáš Mikolov, 2012, word2vec

https://projector.tensorflow.org/

Word embeddings

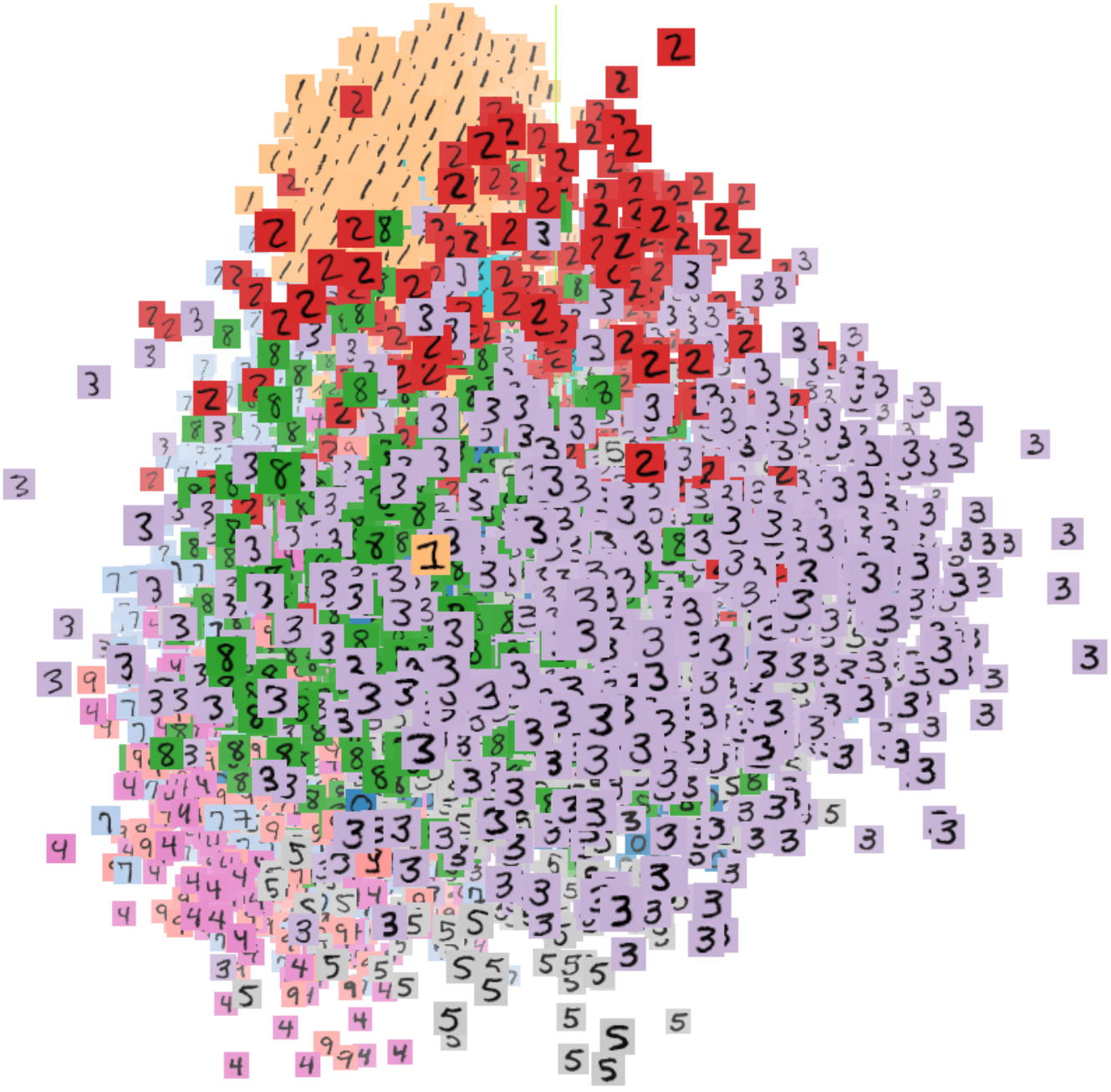


Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese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

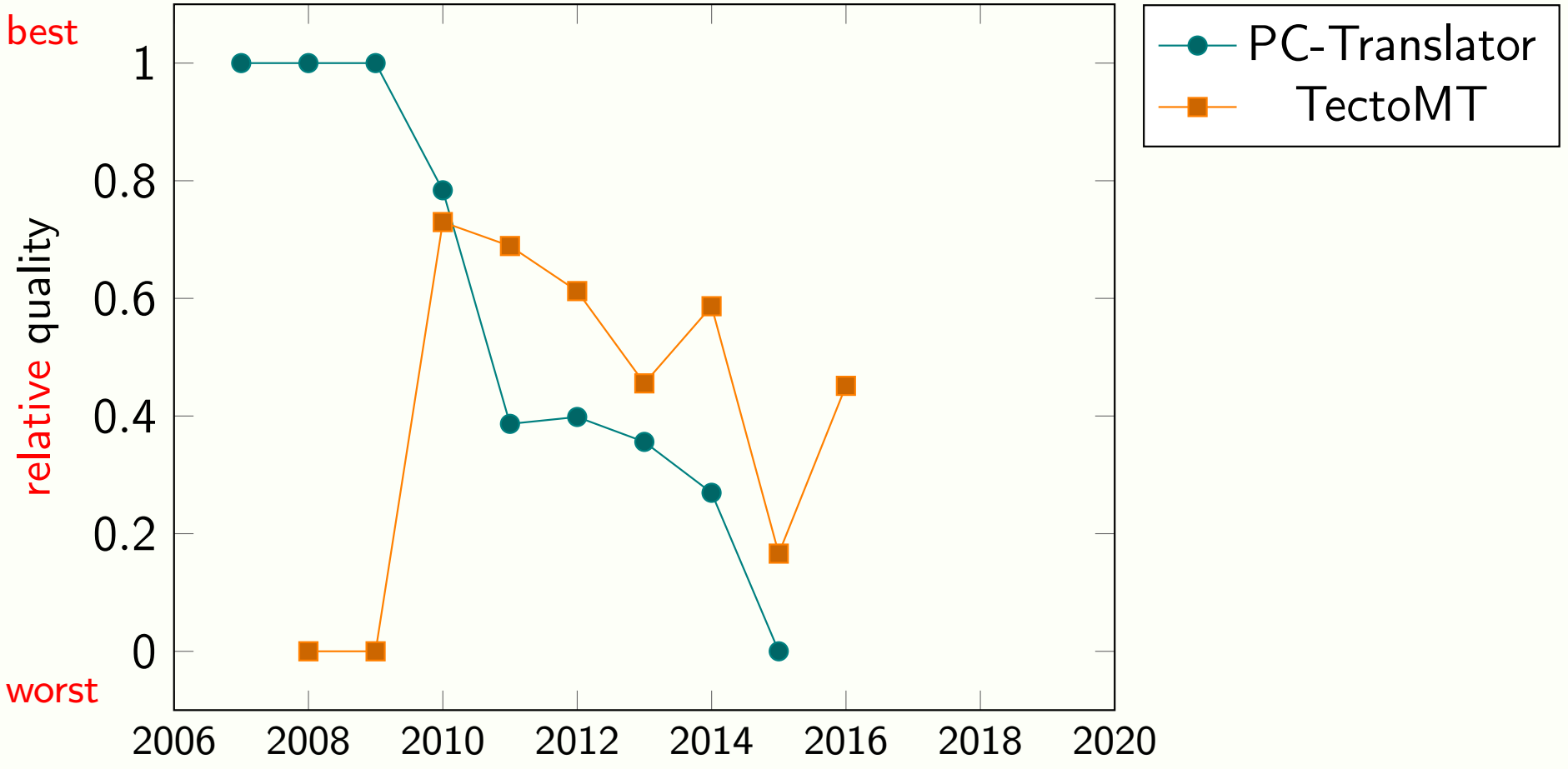
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

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

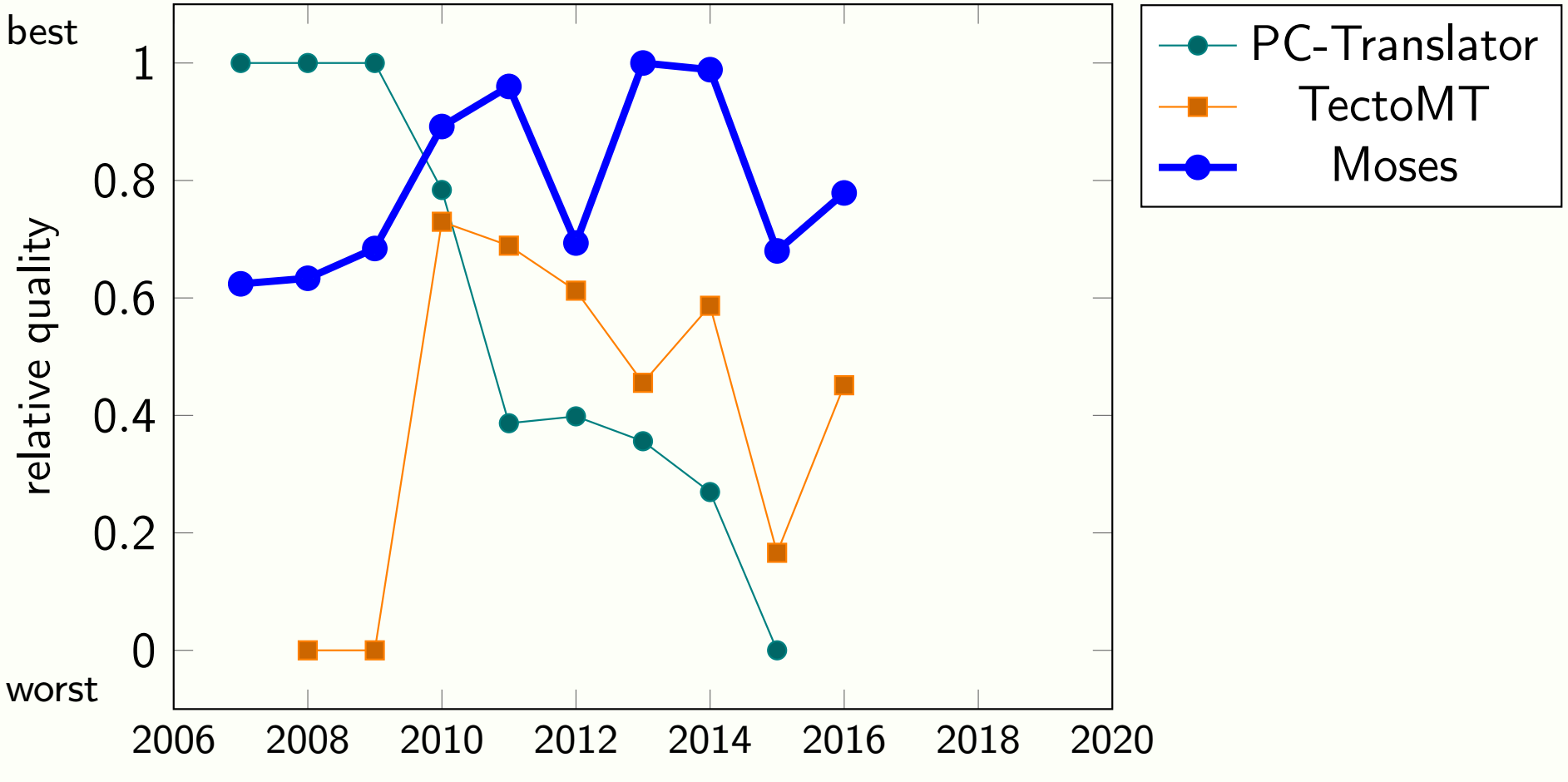
Word embeddings



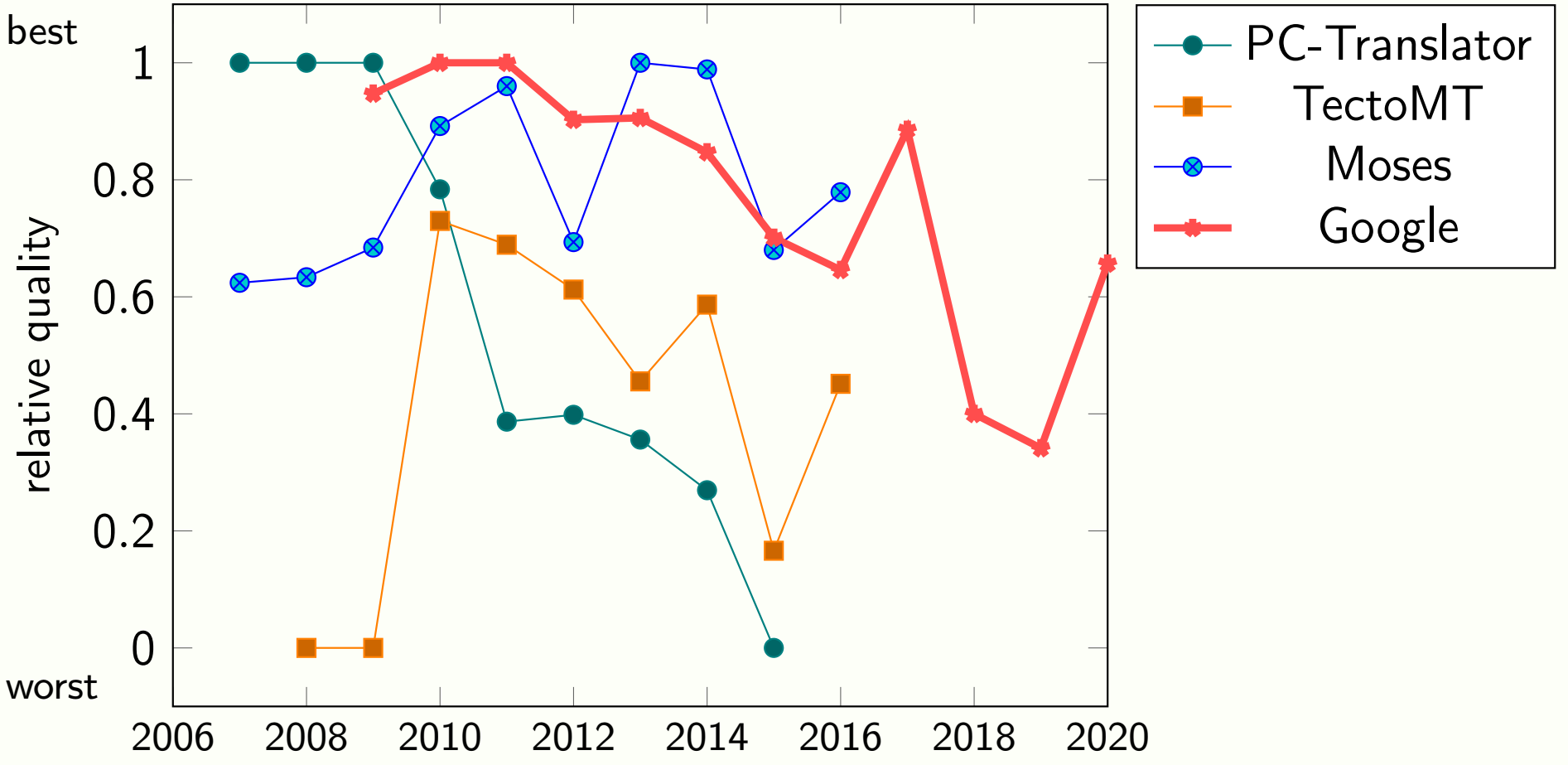
English→Czech MT in 2007–2020



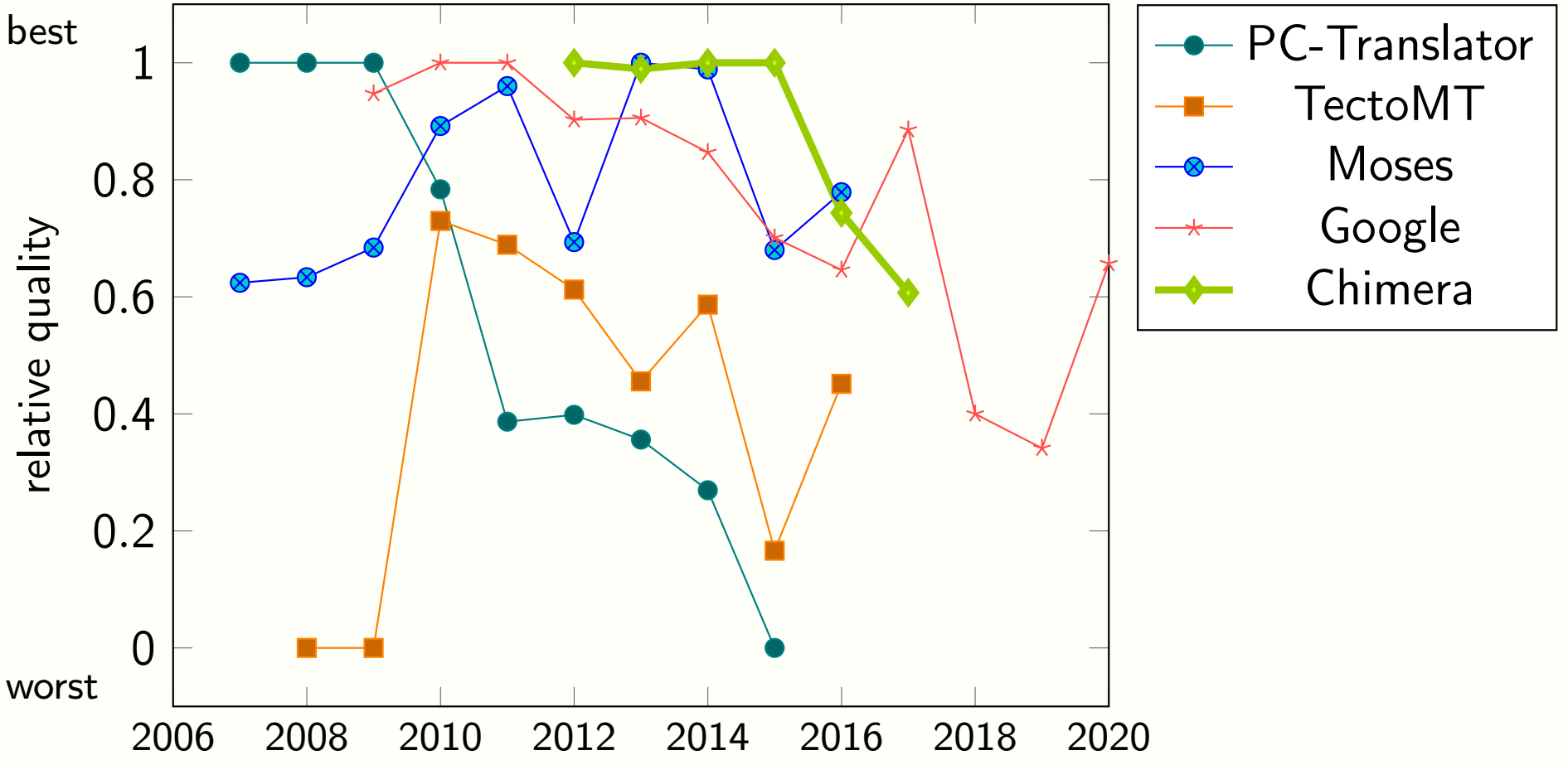
English→Czech MT in 2007–2020



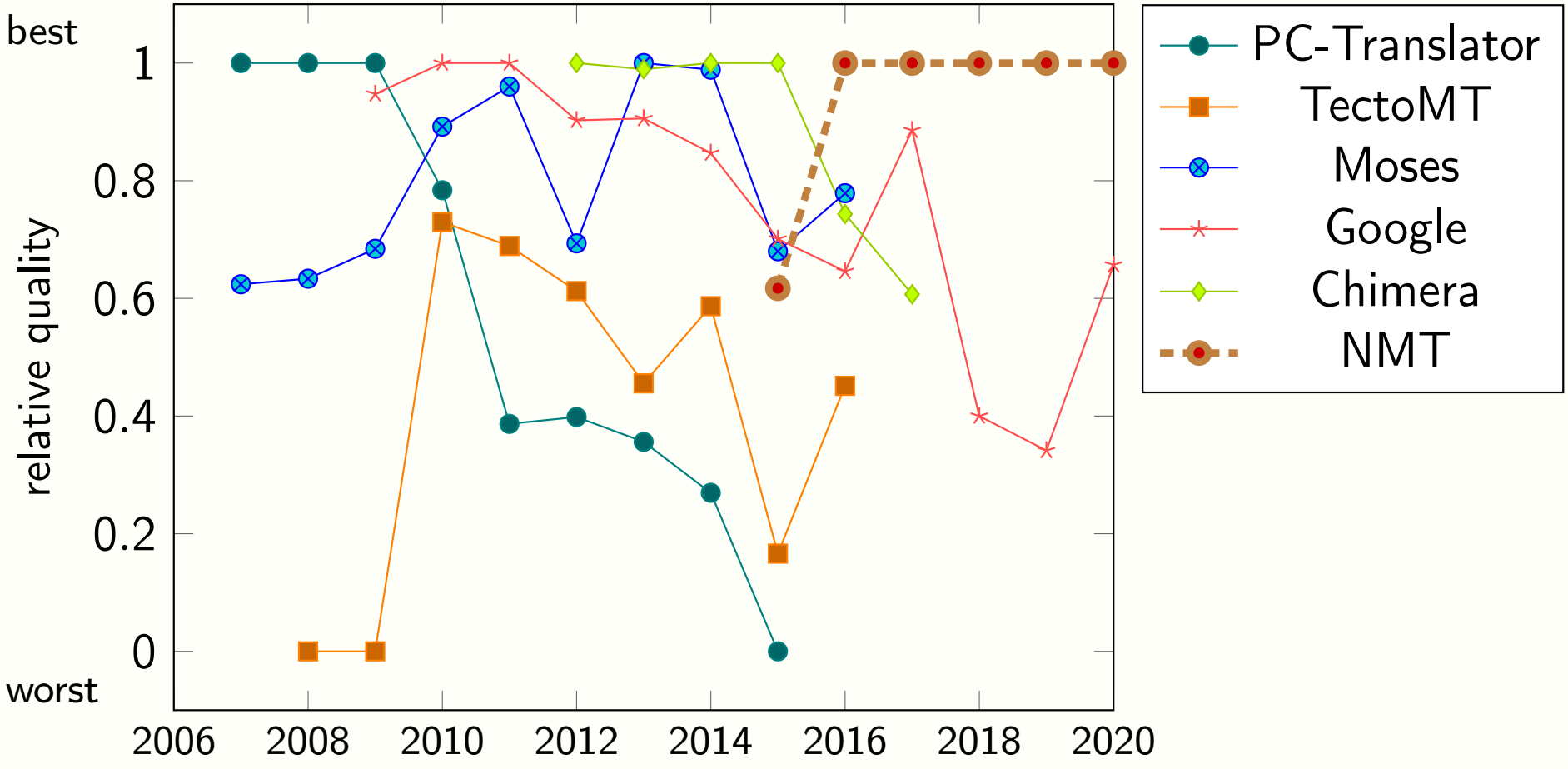
English→Czech MT in 2007–2020



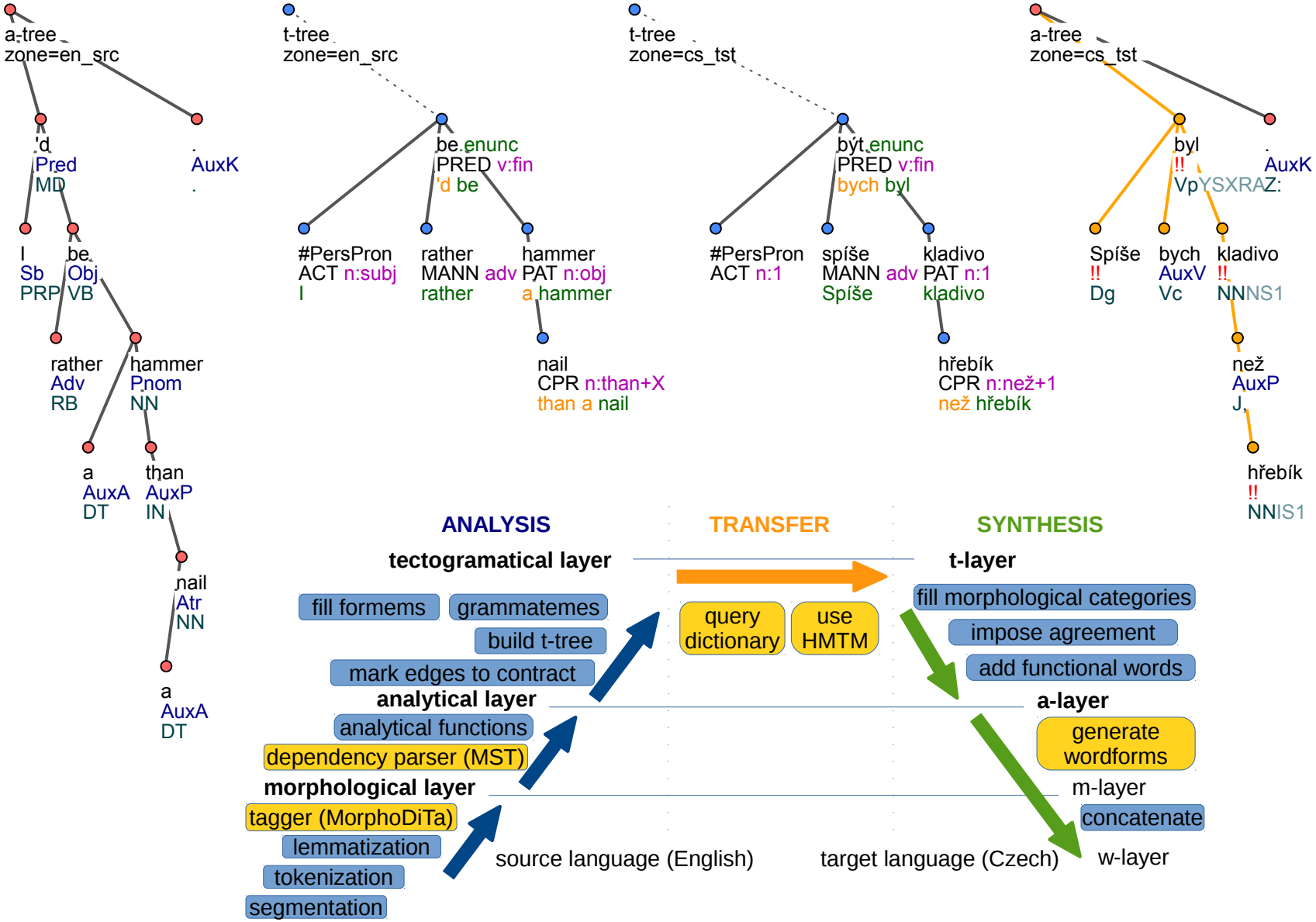
English→Czech MT in 2007–2020



English→Czech MT in 2007–2020



Deep-syntactic translator TectoMT



I'd rather be a hammer than a nail.

Spíše bych byl kladivo než hřebík/nehet.

Machine learning: features in the translation model

output_label=hřebík#N	
feature	λ
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	**0.75**
tense_g=post	0.74
+voice_g=active	**0.66**
prev_lemma=drive	0.66
parent_capitalized=1	0.62
formeme=n:from+X	0.60
+prev_lemma=hammer	**0.59**
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

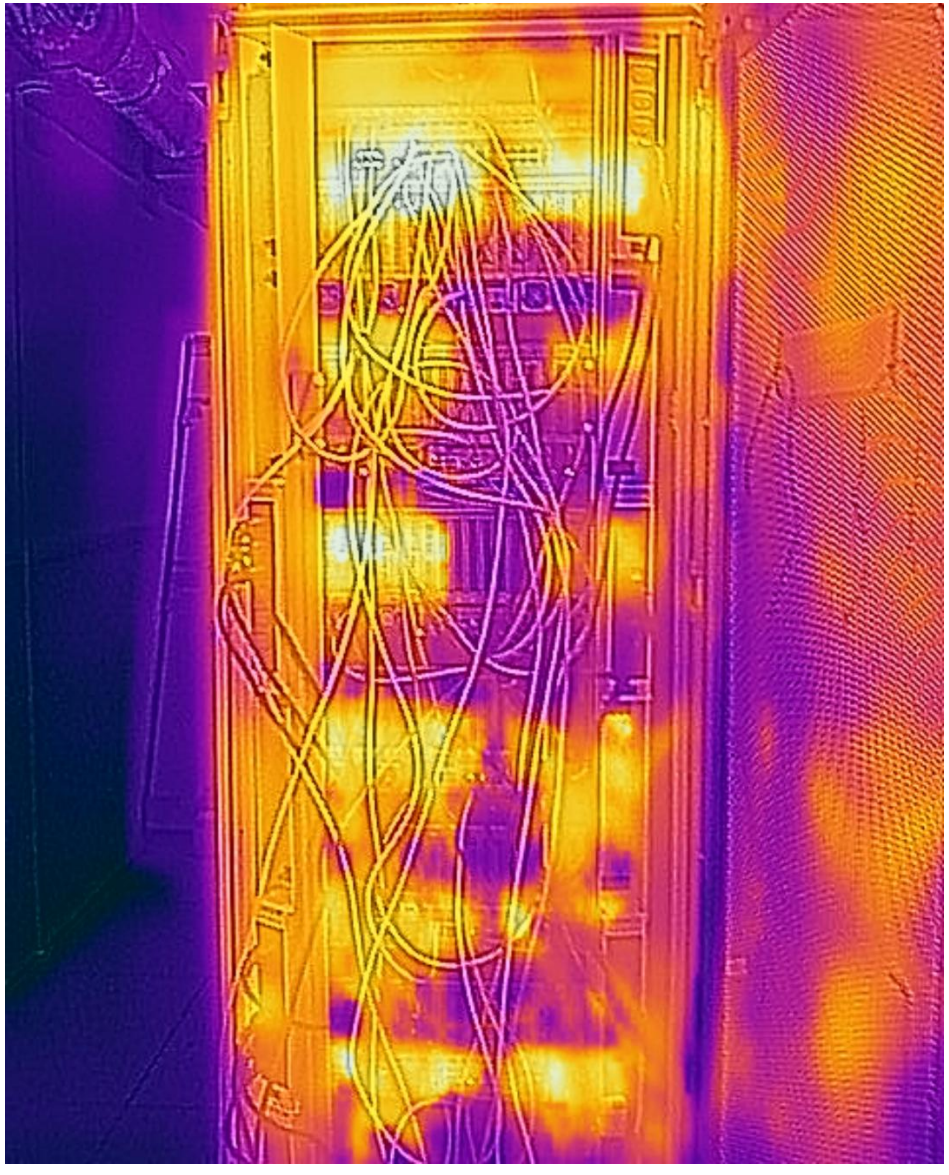
output_label=hřebík#N

feature	λ
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	**0.75**
tense_g=post	0.74
+voice_g=active	**0.66**
prev_lemma=drive	0.66
parent_capitalized=1	0.62
formeme=n:from+X	0.60
+prev_lemma=hammer	**0.59**
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
...	

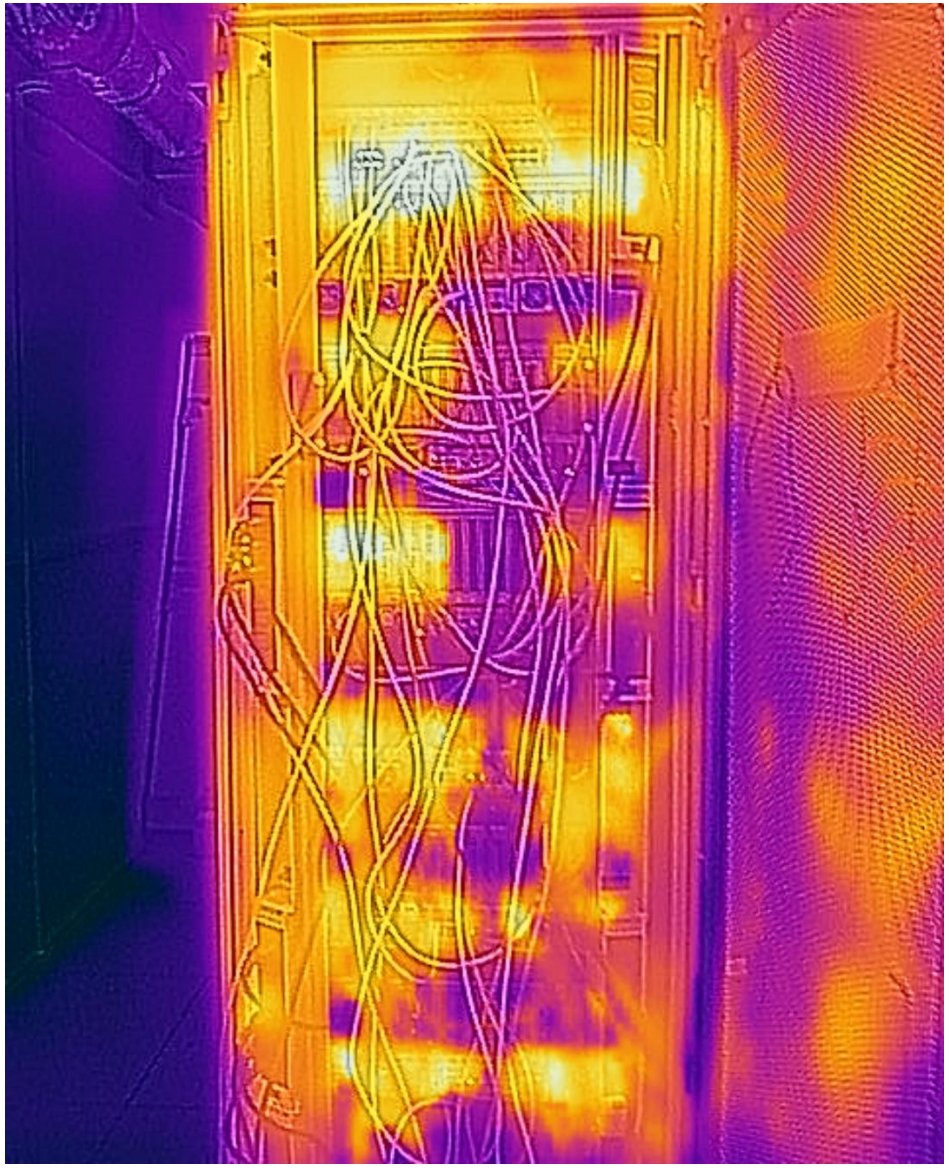
output_label=nehet#N

feature	λ
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	**0.56**
formeme=n:obj	0.53
formeme=n:by+X	0.52
...	

What is this?



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

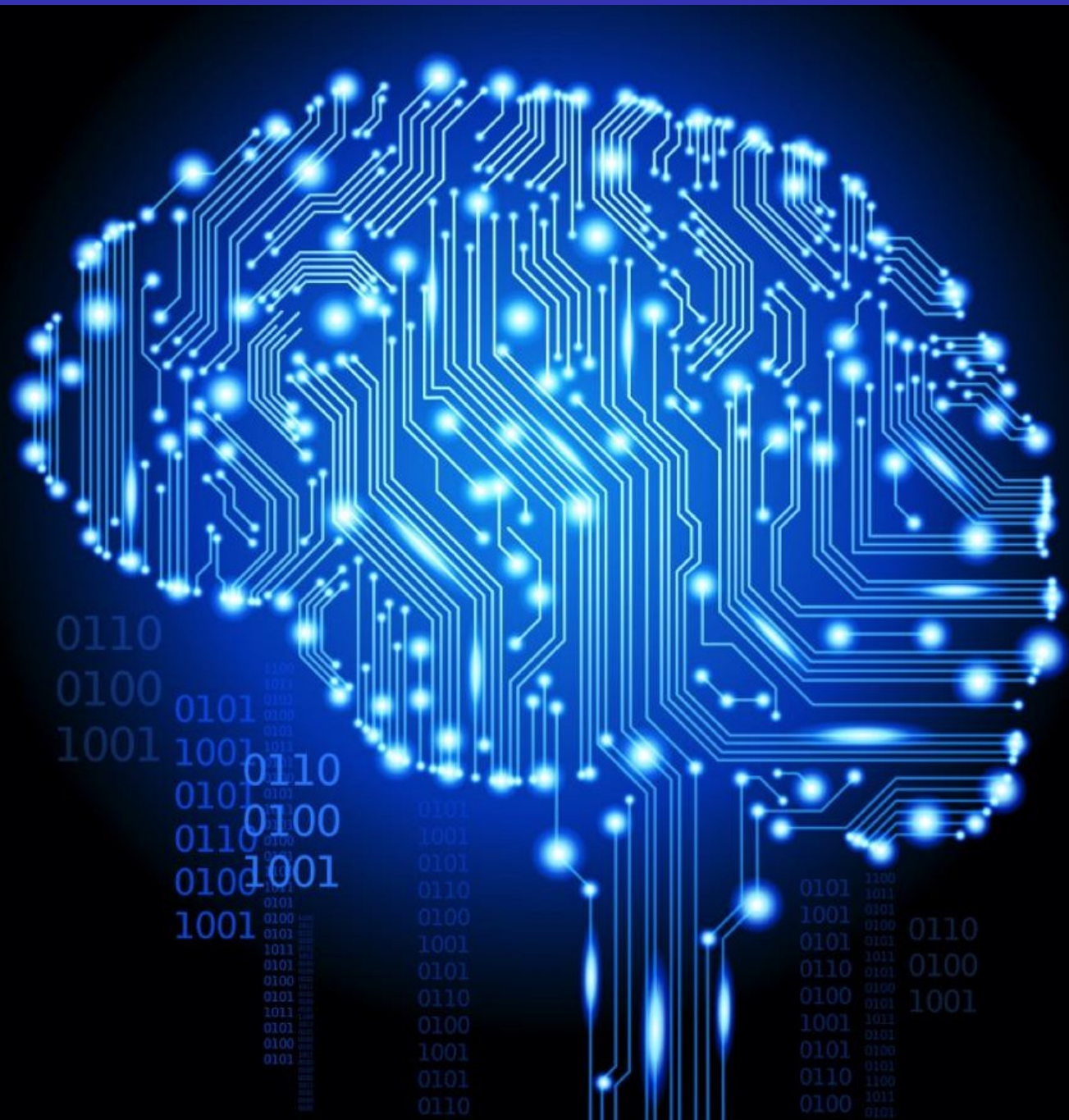


Artificial intelligence and its subfields

artificial intelligence
$\sim$1950

machine learning
$\sim$1980

deep learning
$\sim$2010

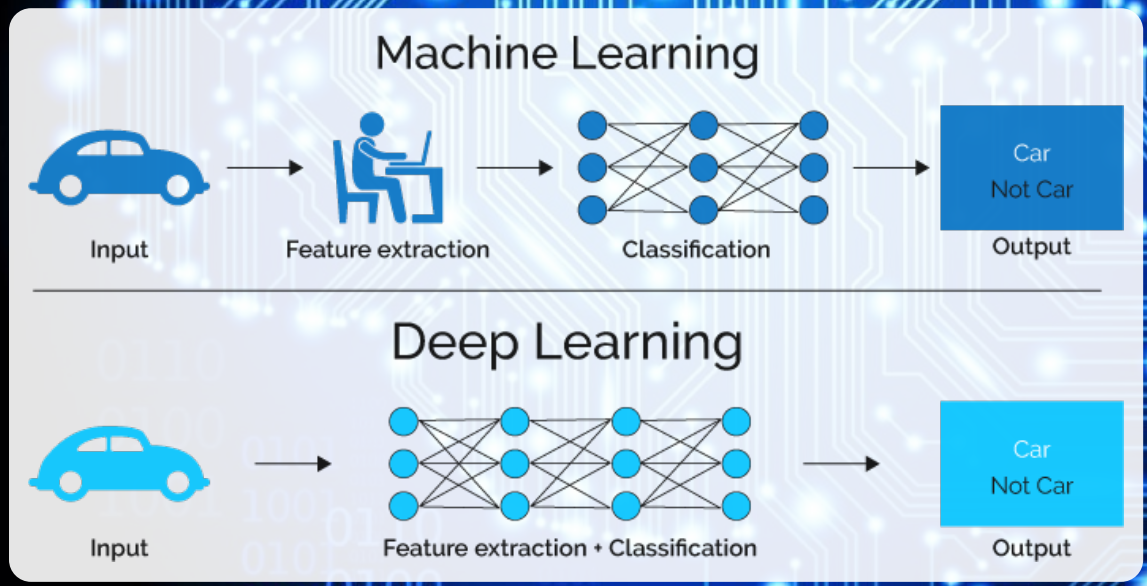


Artificial intelligence and its subfields

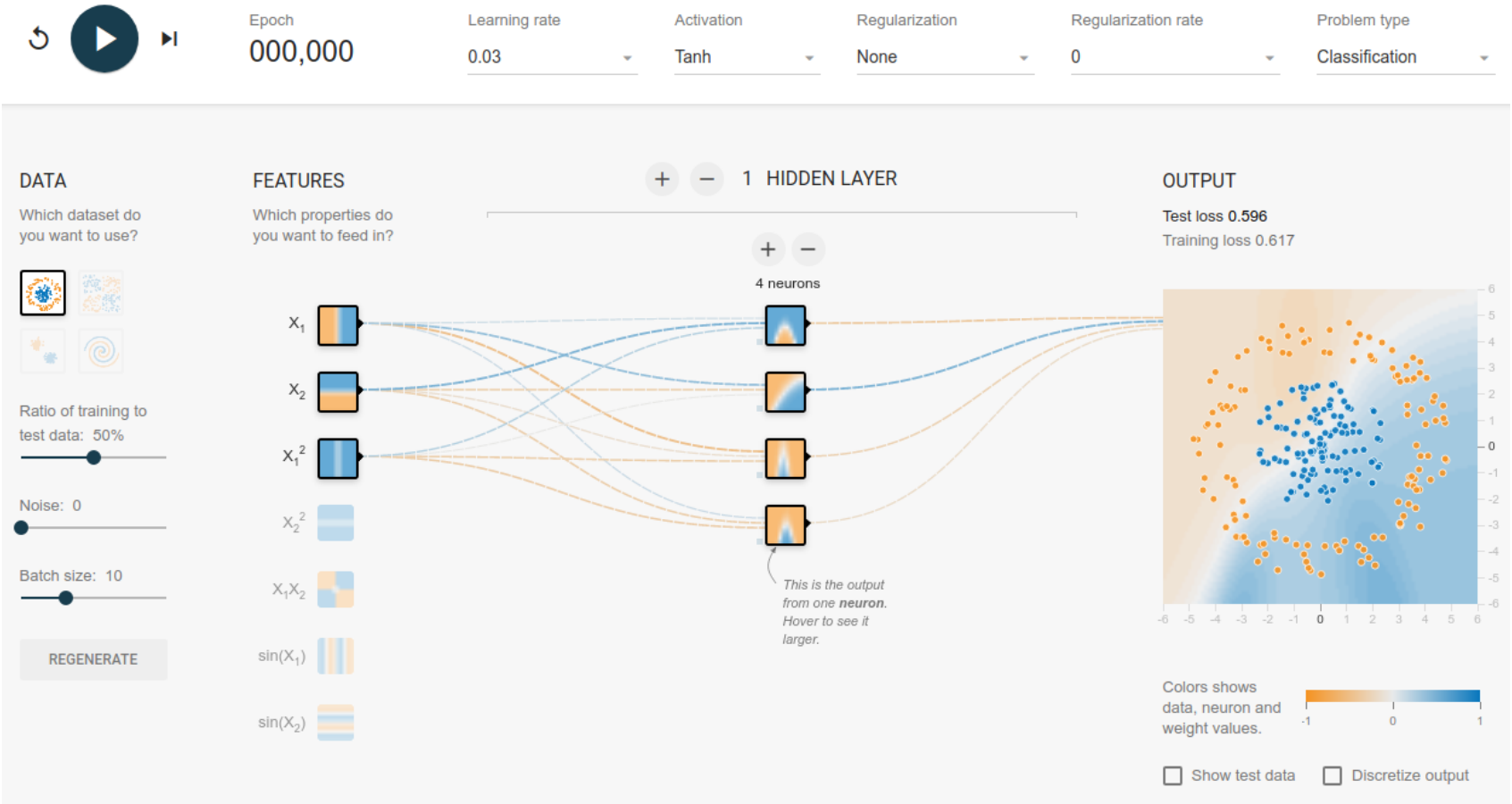
artificial intelligence
~1950

machine learning
~1980

deep learning
~2010

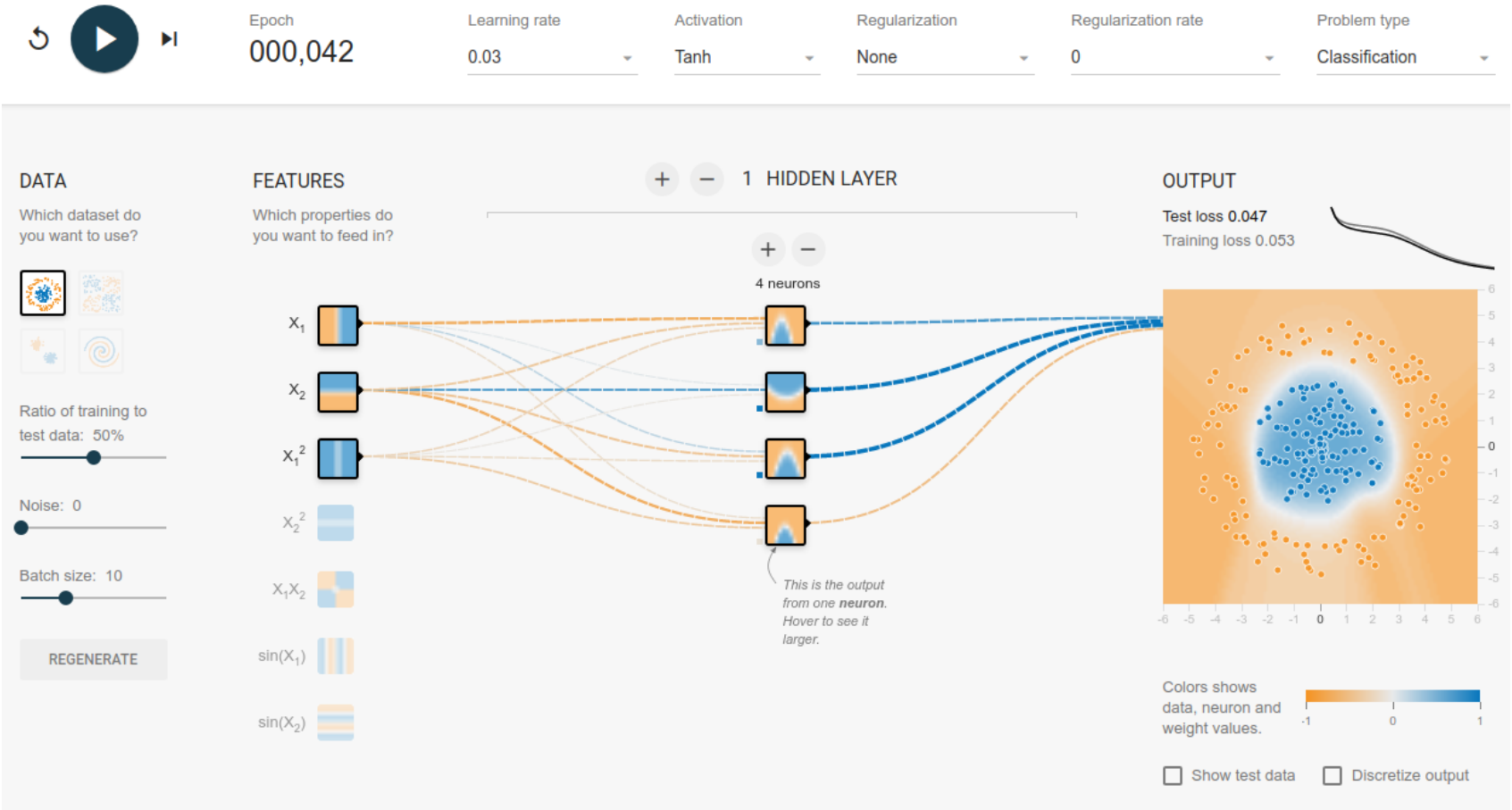


a simple architecture (16 parameters)



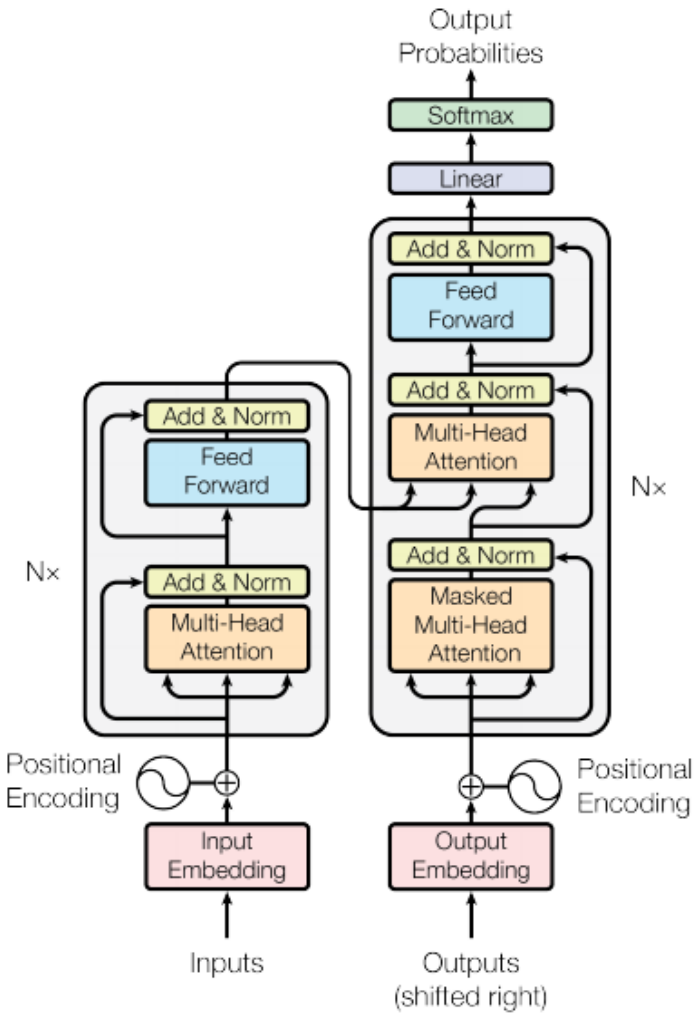
https://playground.tensorflow.org

a simple architecture (16 parameters)

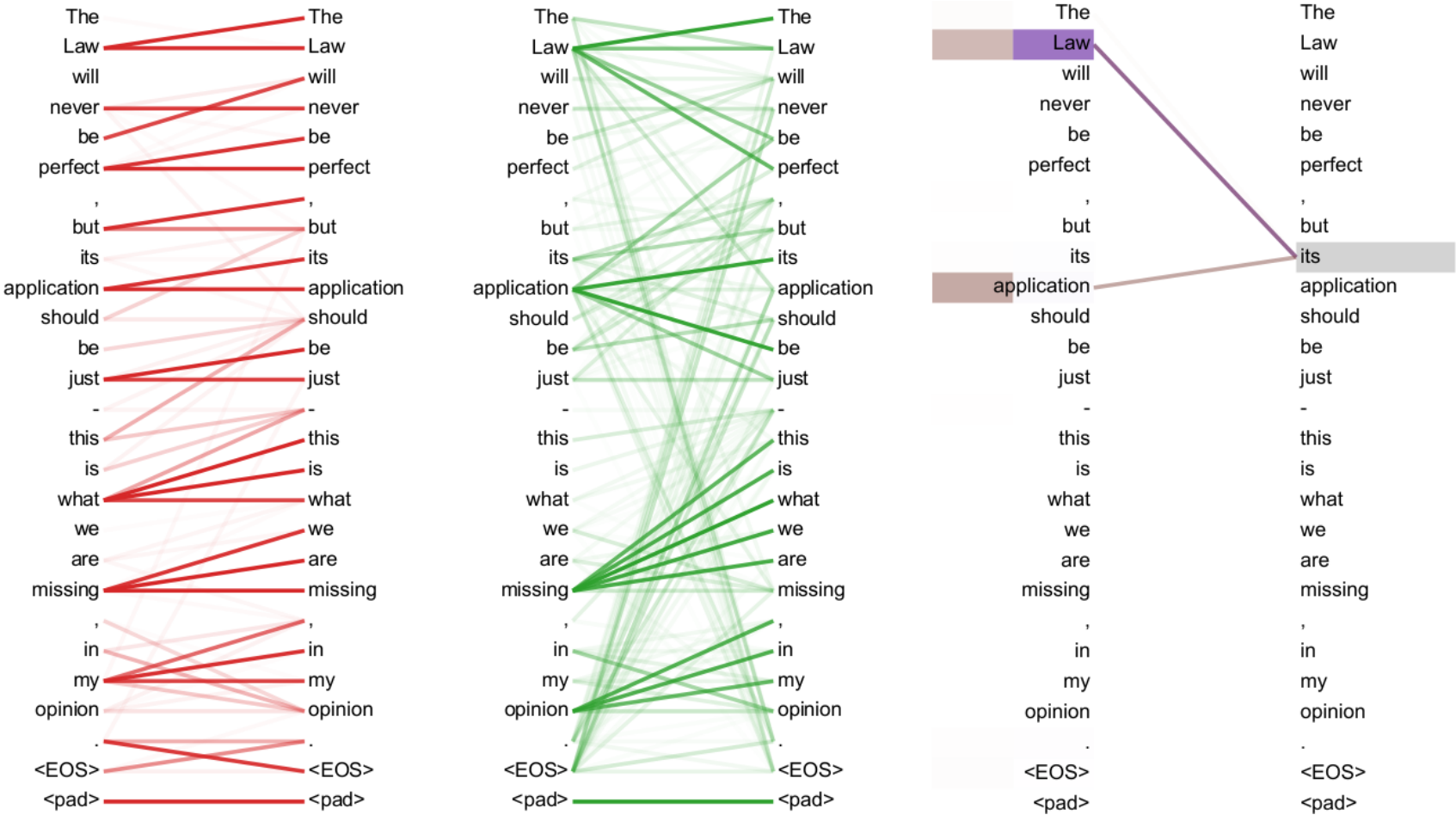


https://playground.tensorflow.org

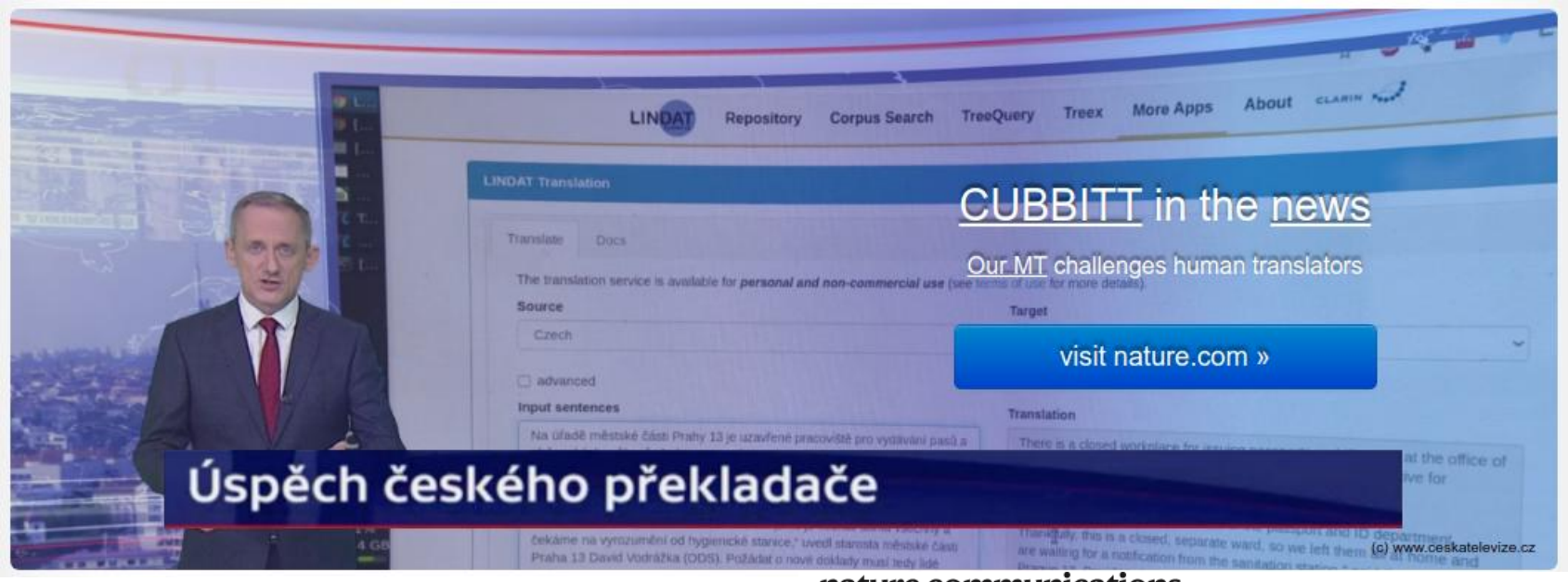
Transformer architecture (213 millions parameters)



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



Why CUBBITT made the headlines in 2020?



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

nature communications

nature › nature communications › articles › article

Article | Open Access | Published: 01 September 2020

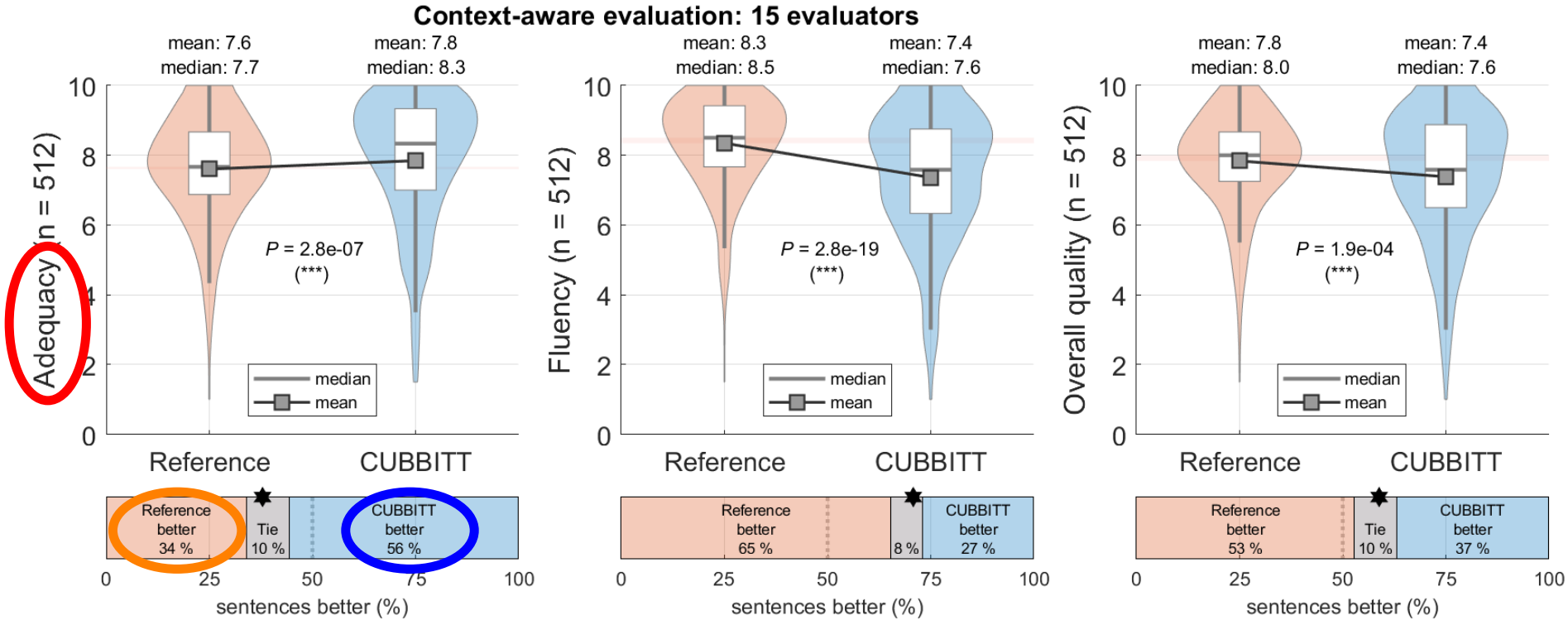
Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals

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

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

The main result of our human evaluation

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



MT can be dangerous

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)

MT can be dangerous

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

Document-level translation

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. **She** works as a conductor.

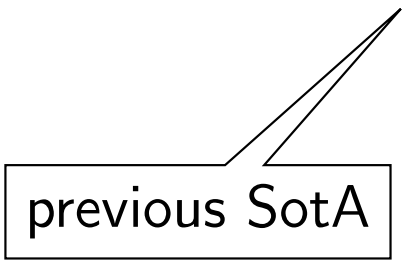
Source	Saša je muž. Pracuje jako průvodčí.
DeepL	Sasha is a man. She works as a conductor.
CUBBITT-doc	Sasha is a man. **He** works as a conductor.

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?

Backtranslation (Sennrich et al., 2016)

- 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

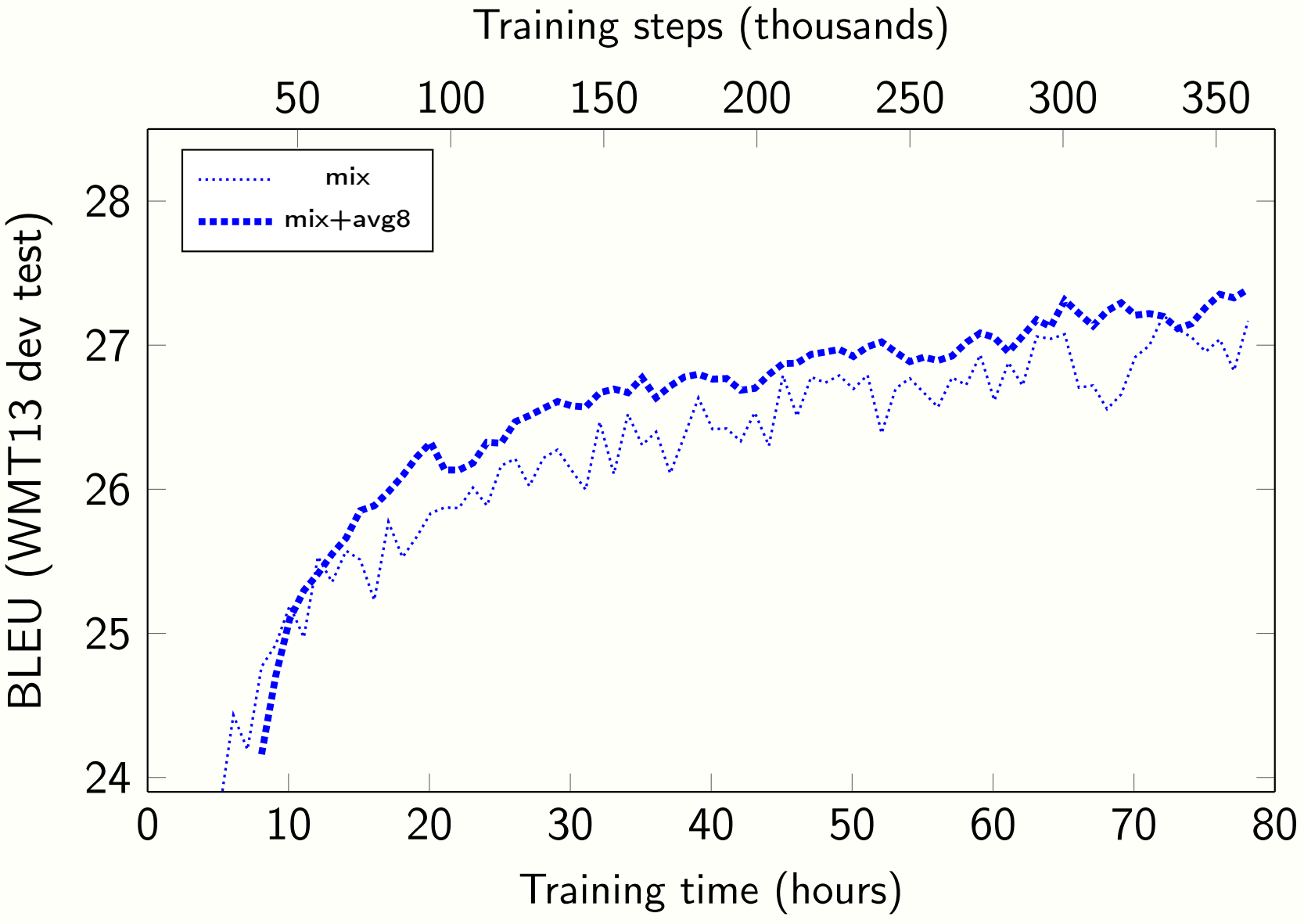


Backtranslation (Sennrich et al., 2016)

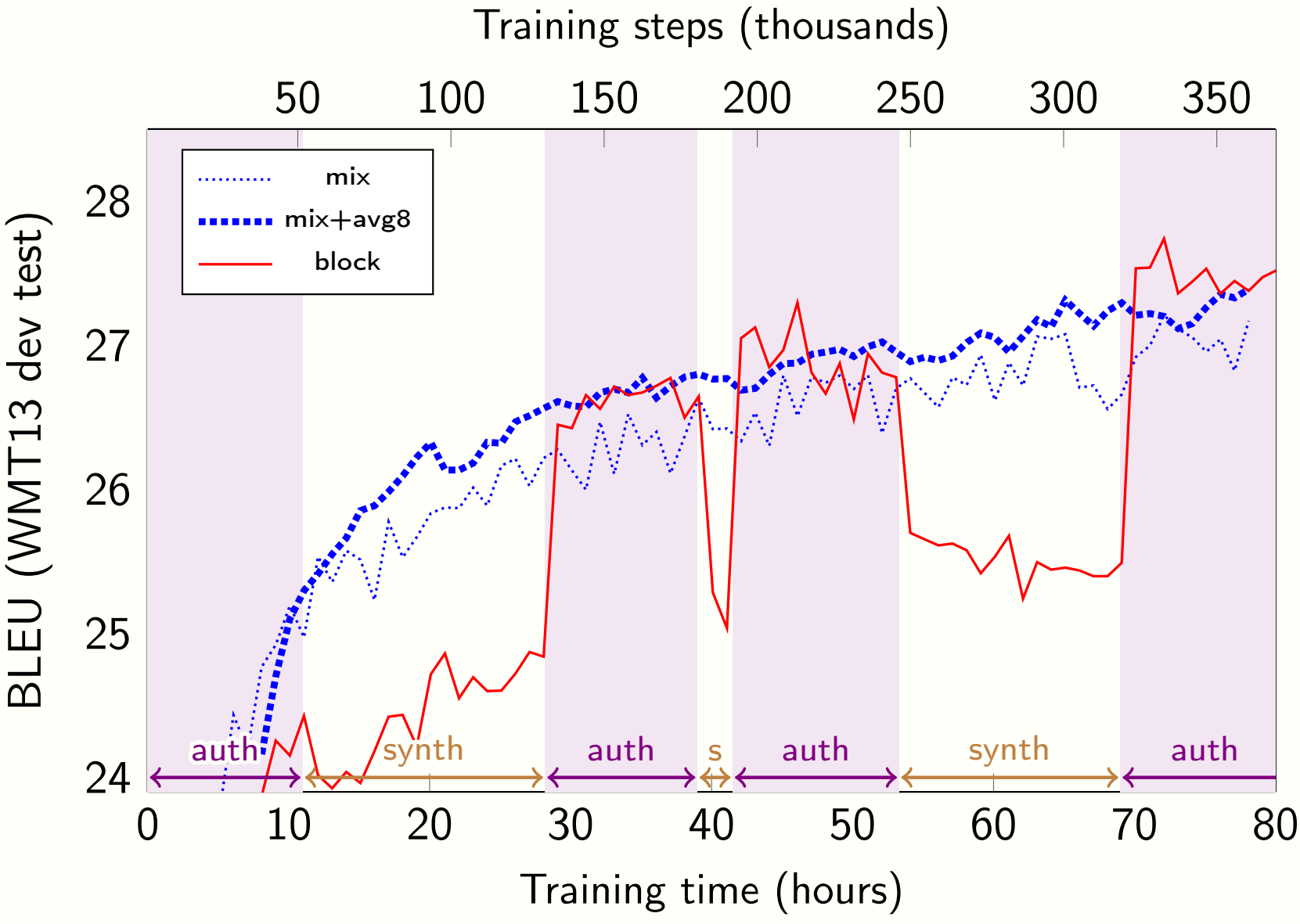
- For EN→CS translation, we can exploit monolingual CS data.
- Translate the data back to English (with any CS→EN MT).
- Prepare synthetic parallel data (orig-CS, synth-EN).
- Train on both authentic and synthetic
  - **fine-tune BT**: first auth then auth+synth
  - **mix BT**: shuffle auth and synth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

Block Backtranslation

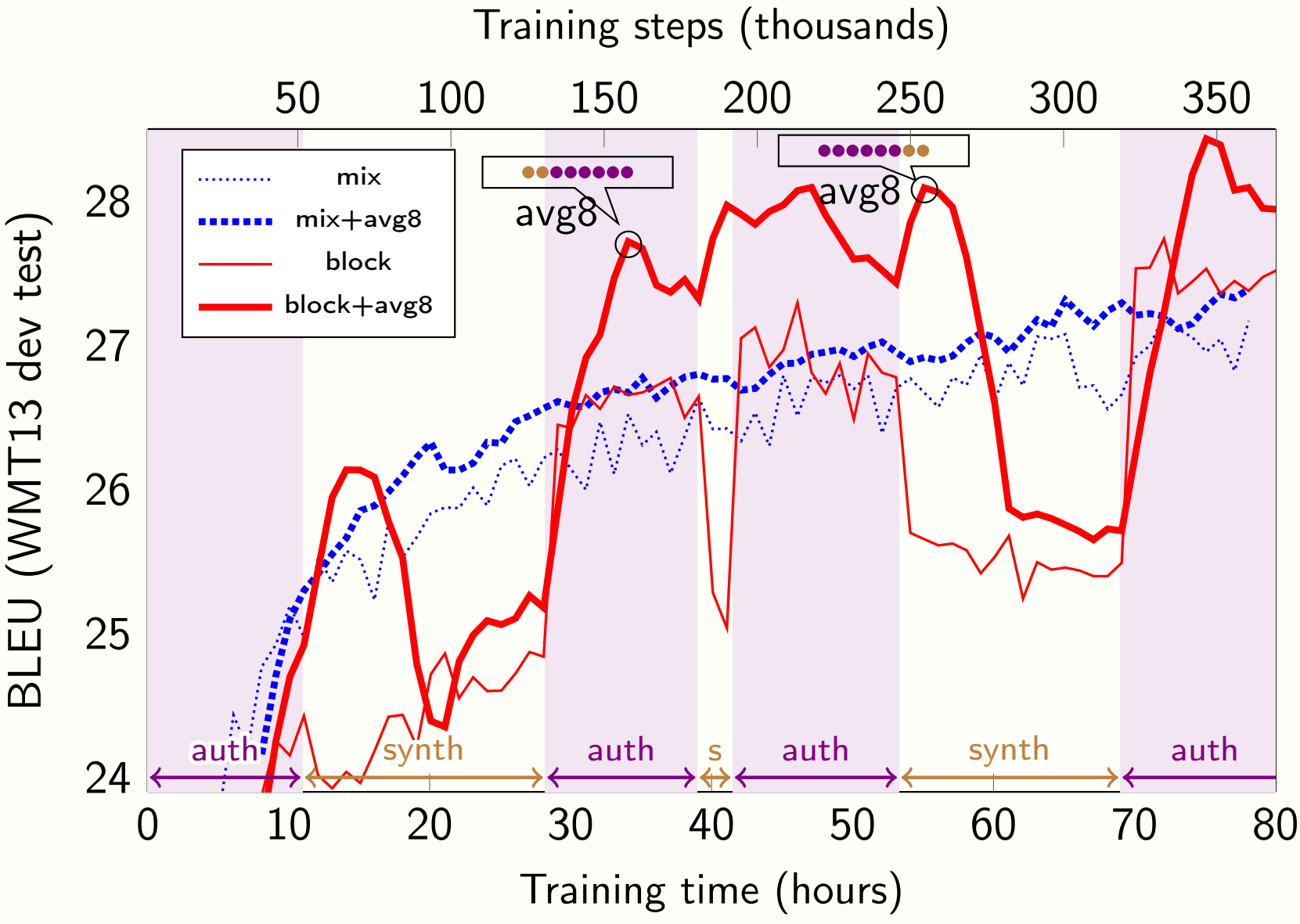
Block Backtranslation



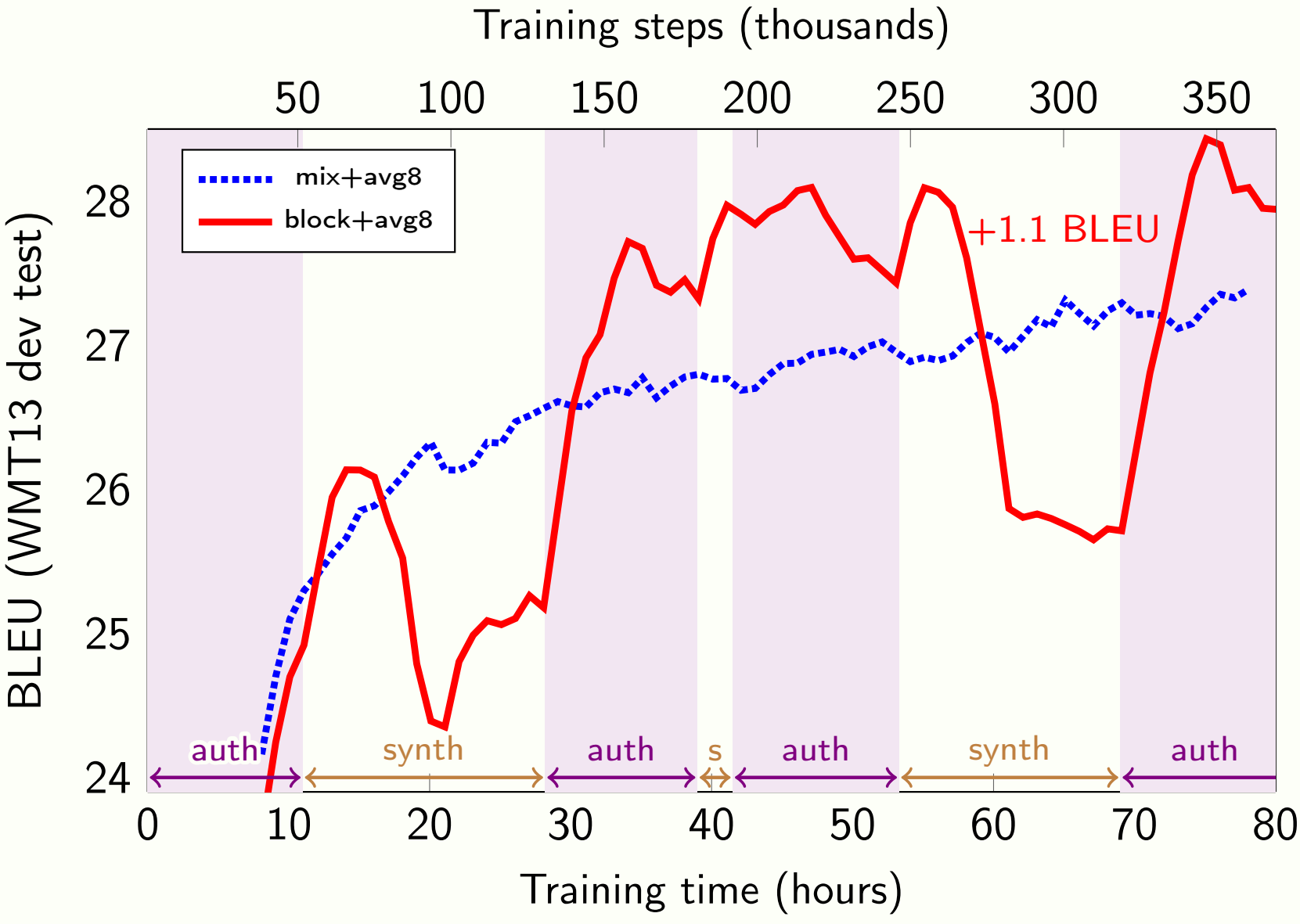
Block Backtranslation



Block Backtranslation



Block Backtranslation

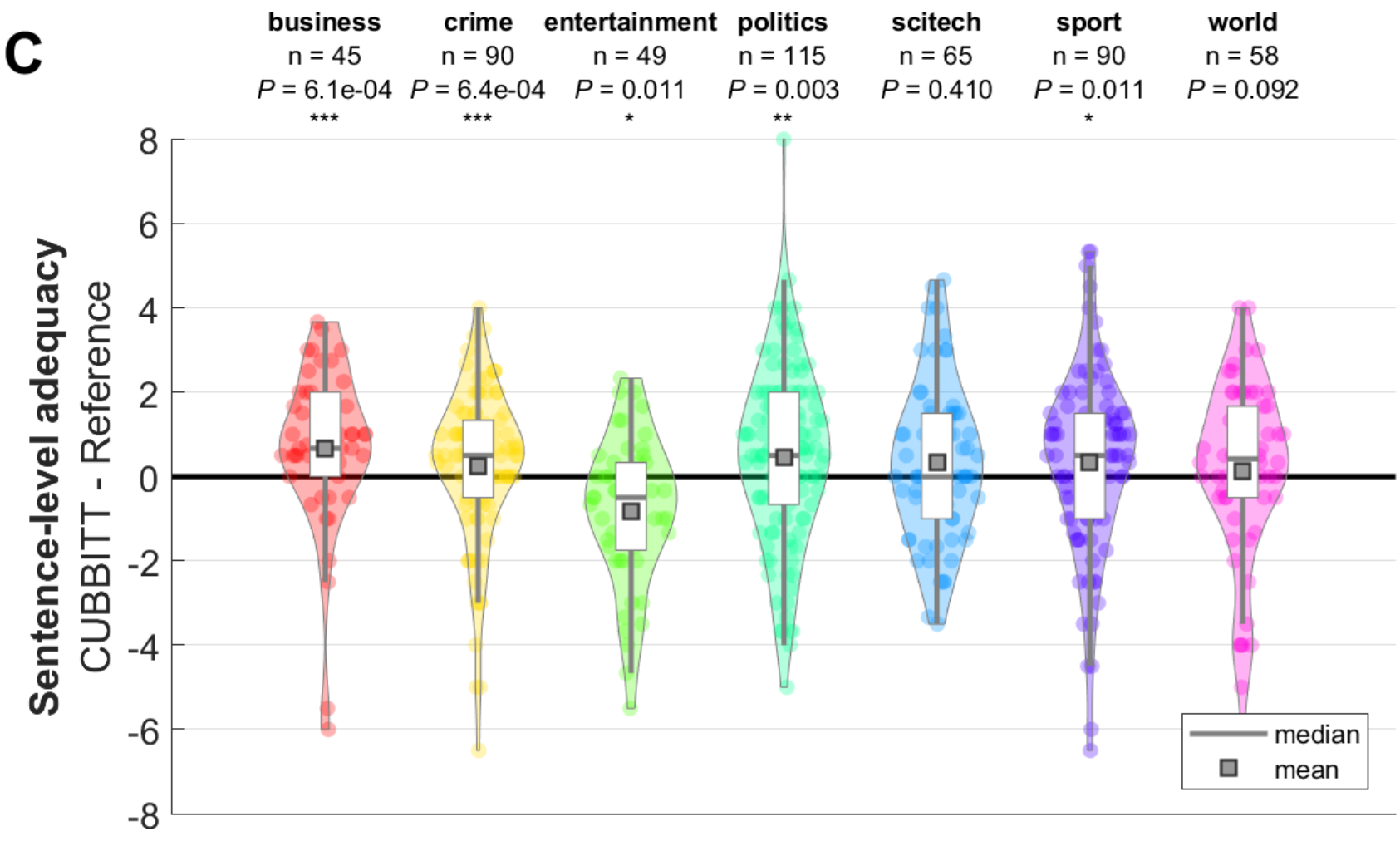


Manual evaluation example

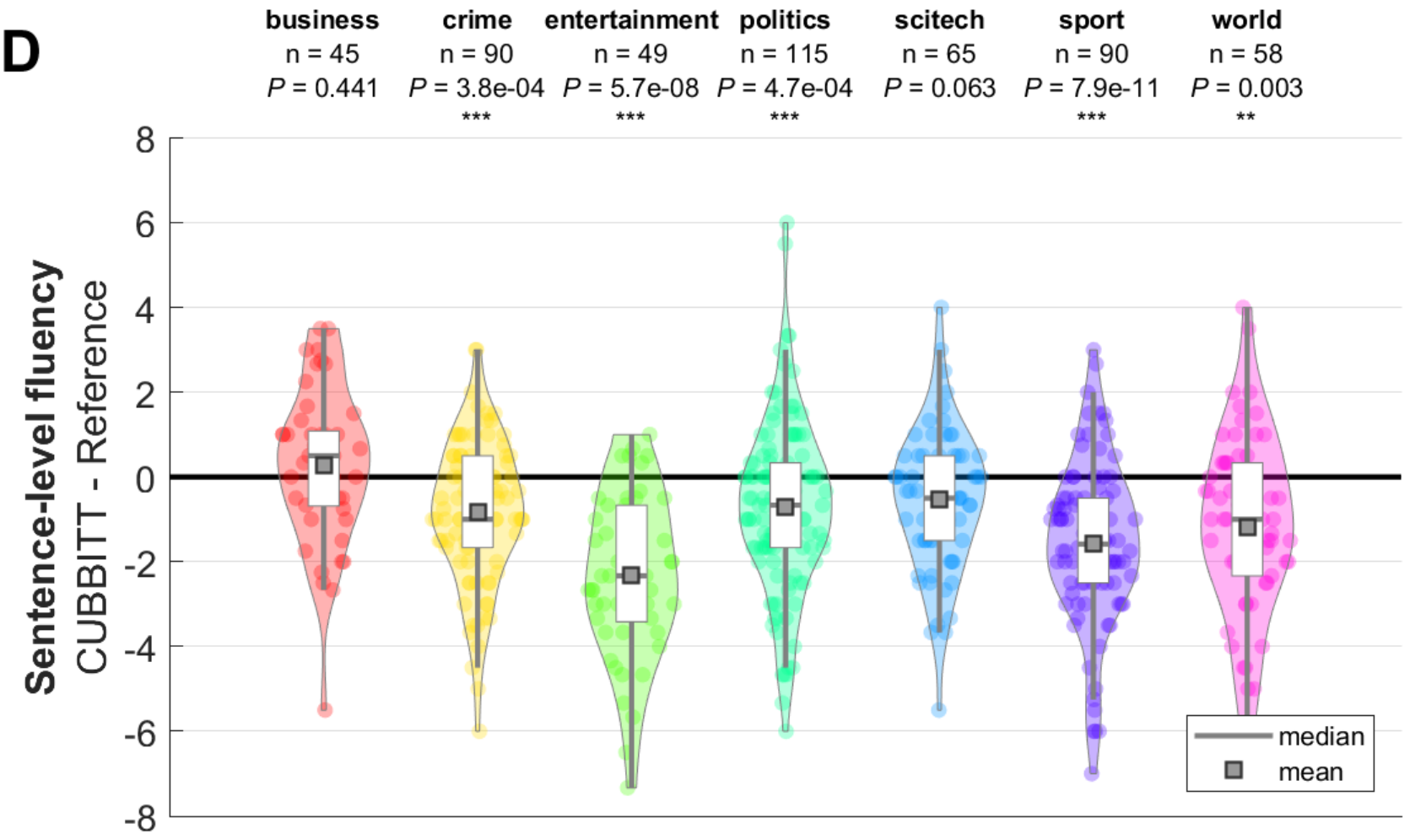
blind, sentence-level but document-aware, side-by-side (RankME)

	G	H	I	J	K	L	M	N	O	P
1	Source	Translation1	T1_overall	T1_adequacy	T1_fluency	Translation2	T2_overall	T2_adequacy	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, Quinlan 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 Convercent 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 schvalují odvážní kapitáloví investoři, jenž podpořili oba startupy.	7	6	7	Je to obchodní model a příležitost, která má souhlas investorů rizikového kapitálu, kteří podpořili oba start-upy.	10	10	10	T1: chybný překlad termínu "venture capital"
174	Convercent raised $10 million in funding in February from firms such as Sapphire Ventures and Tola Capital, bringing its total capital raised to $47 million.	Convercent vybral v rámci své únorové kampaně od firem jako Sapphire Ventures a Tola Capital celkově 10 milionů $ a nakonec si odnesl kapitál ve výši 47 milionů $.	3	4	3	Convercent 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 společností, kultury a rizikovému kapitálu.	3	4	3	Tyto investice nejsou vzhledem k propojenosti společností, kultury a rizikového kapitálu žádným 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	8	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, nejsou 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 termínu "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



It depends on the input text domain

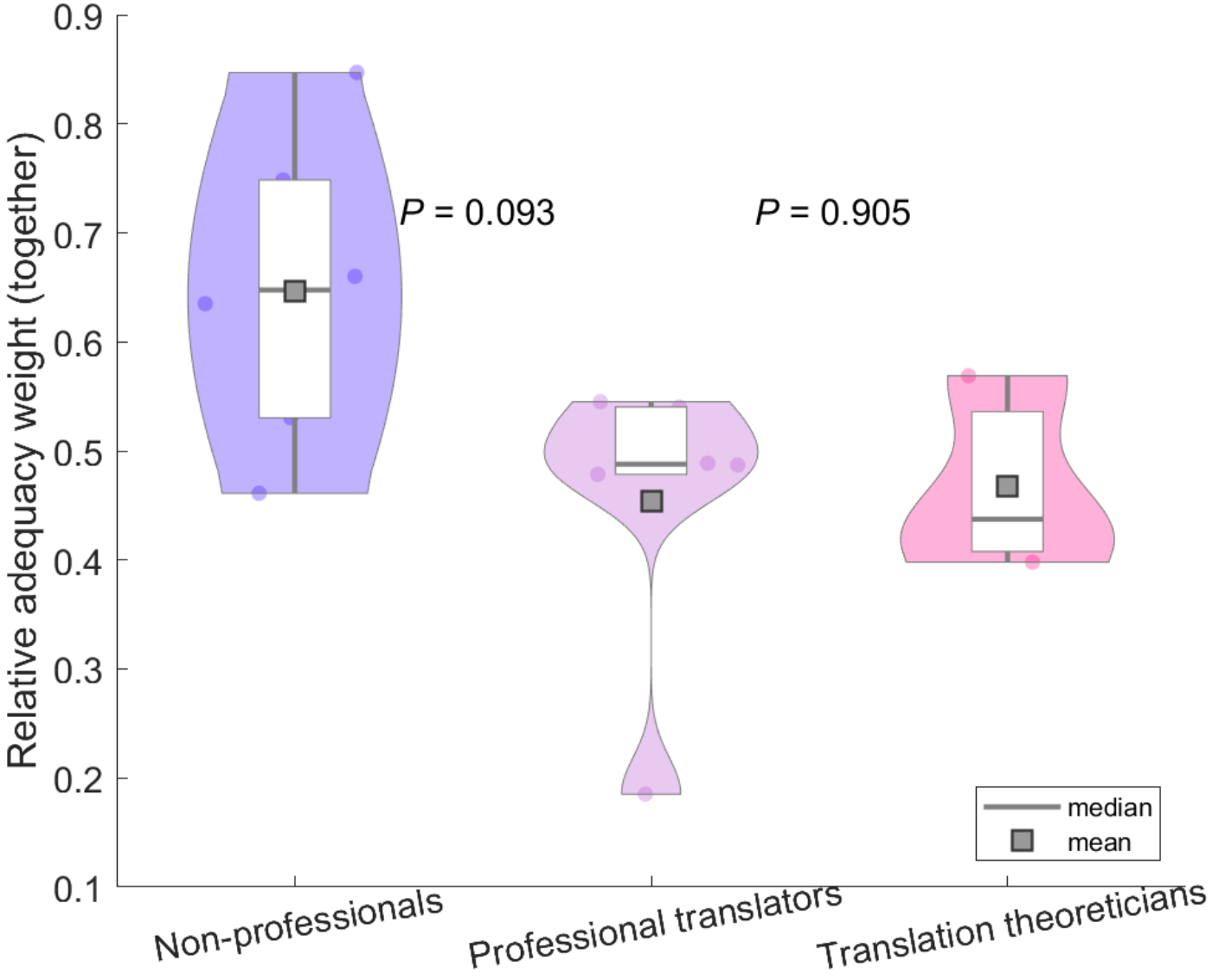


Are three measures necessary?

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

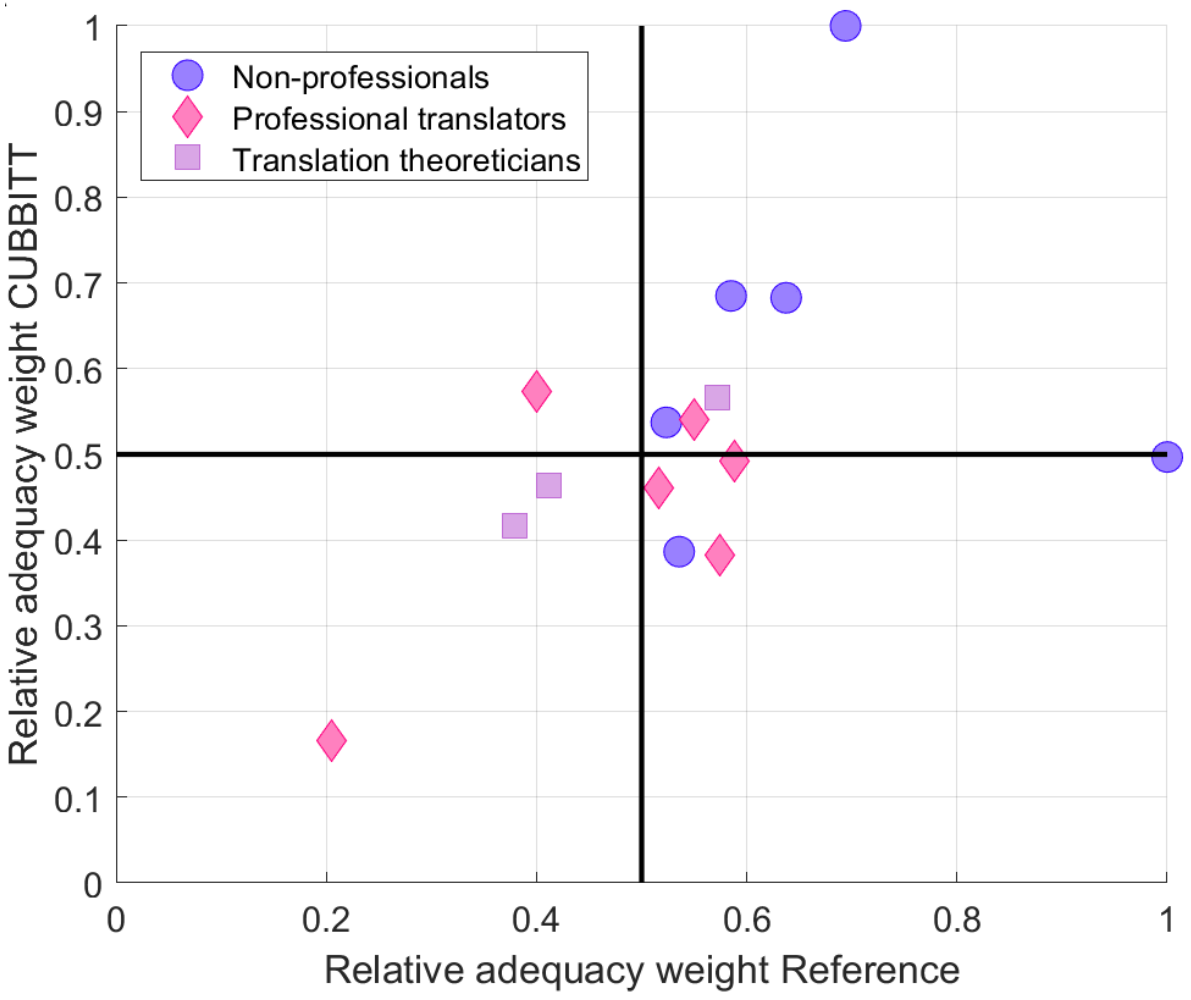
It depends on the annotator

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



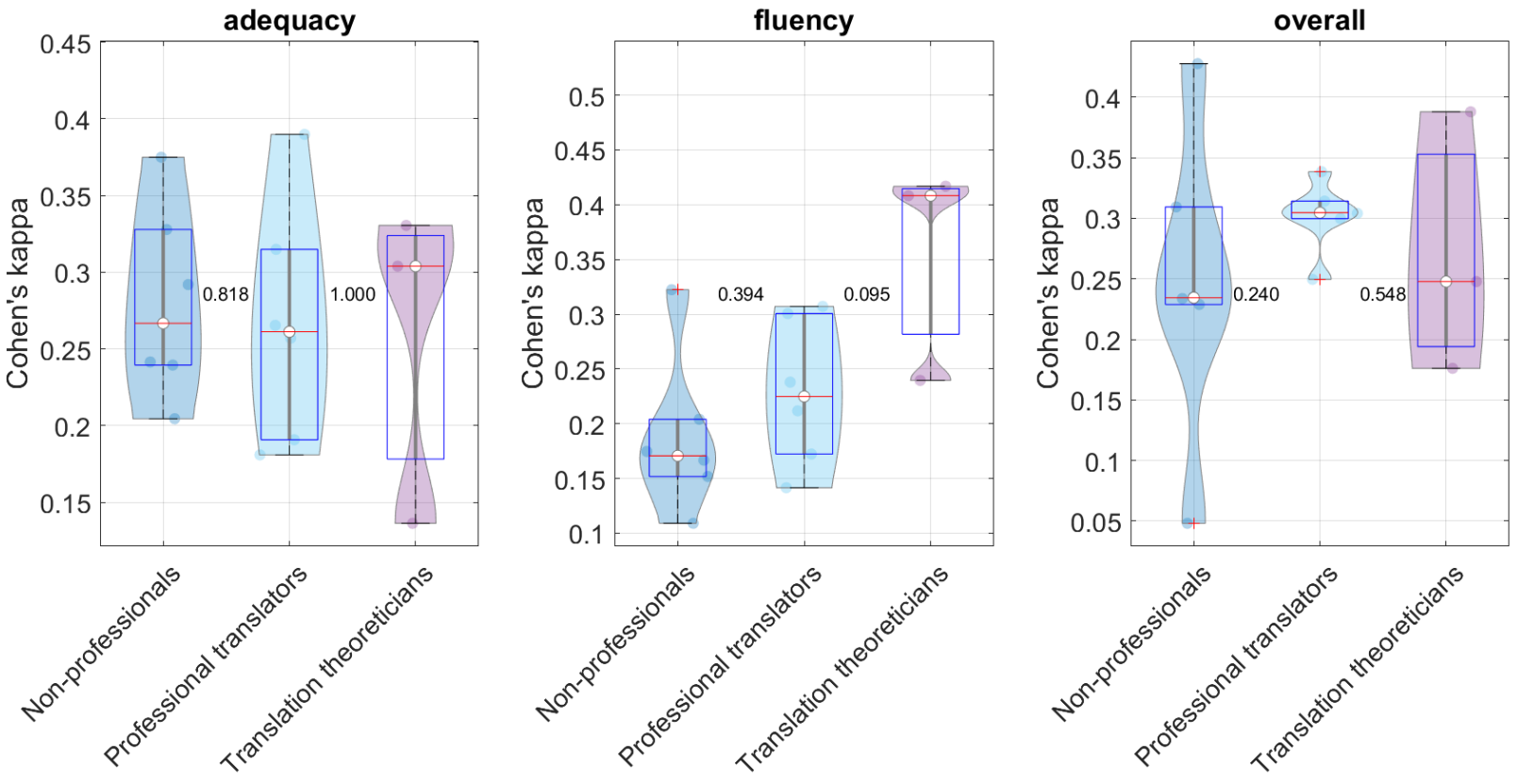
It depends on the annotator

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



Inter-annotator agreement

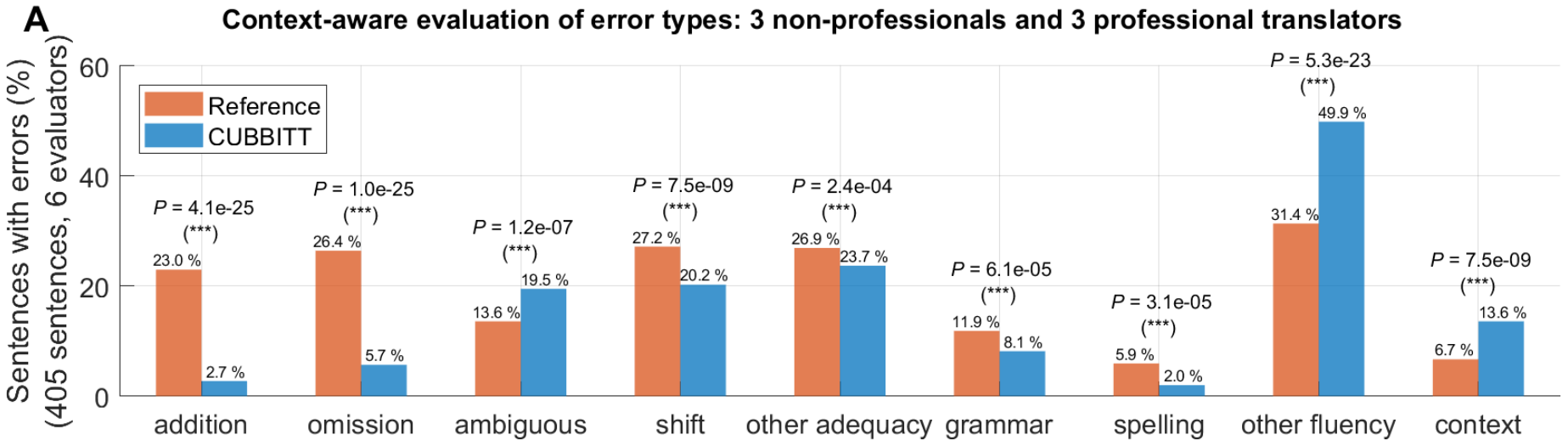
We cannot conclude that professionals are more reliable.



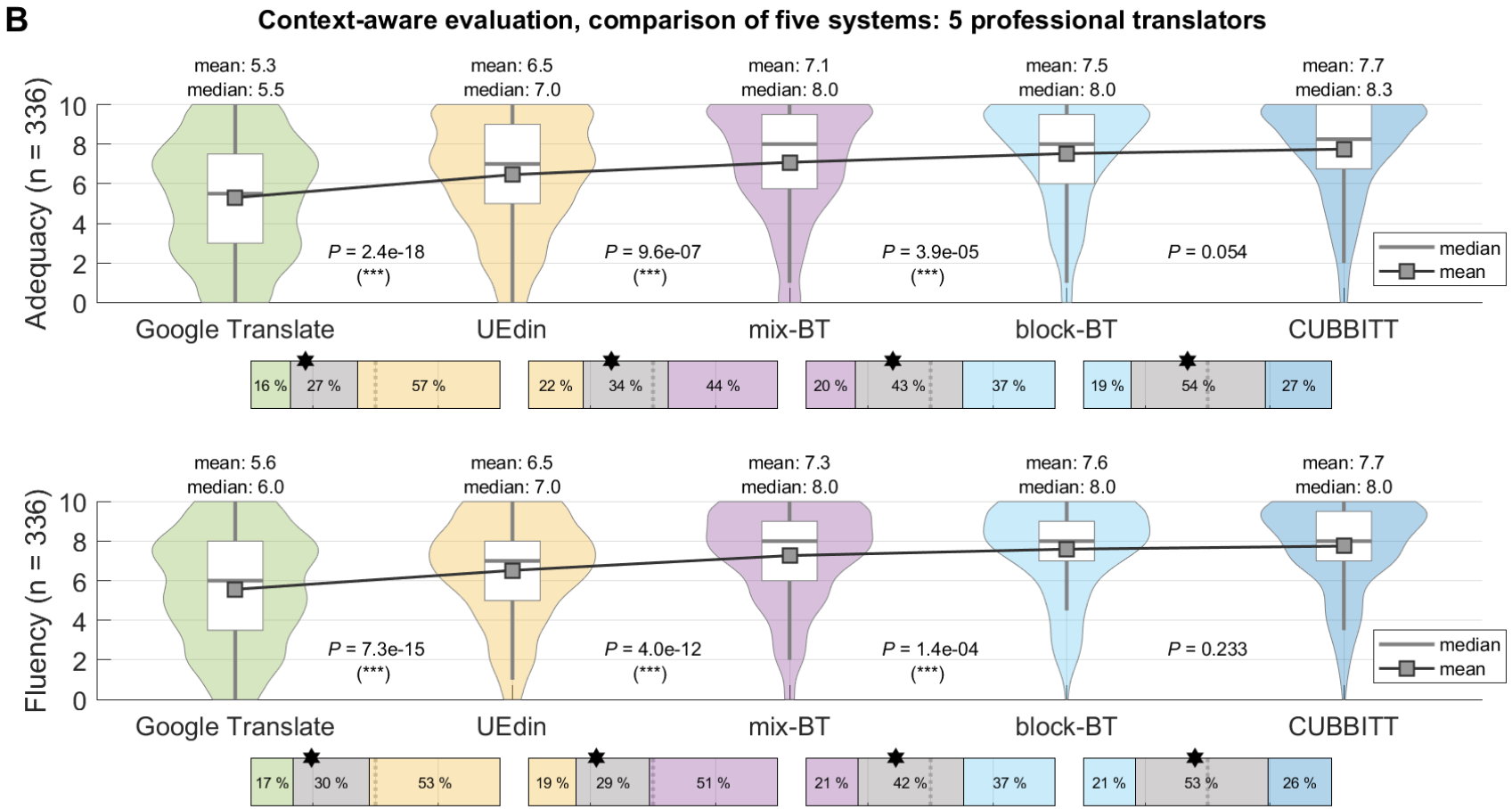
Distribution of error types

CUBBITT makes more errors than humans in

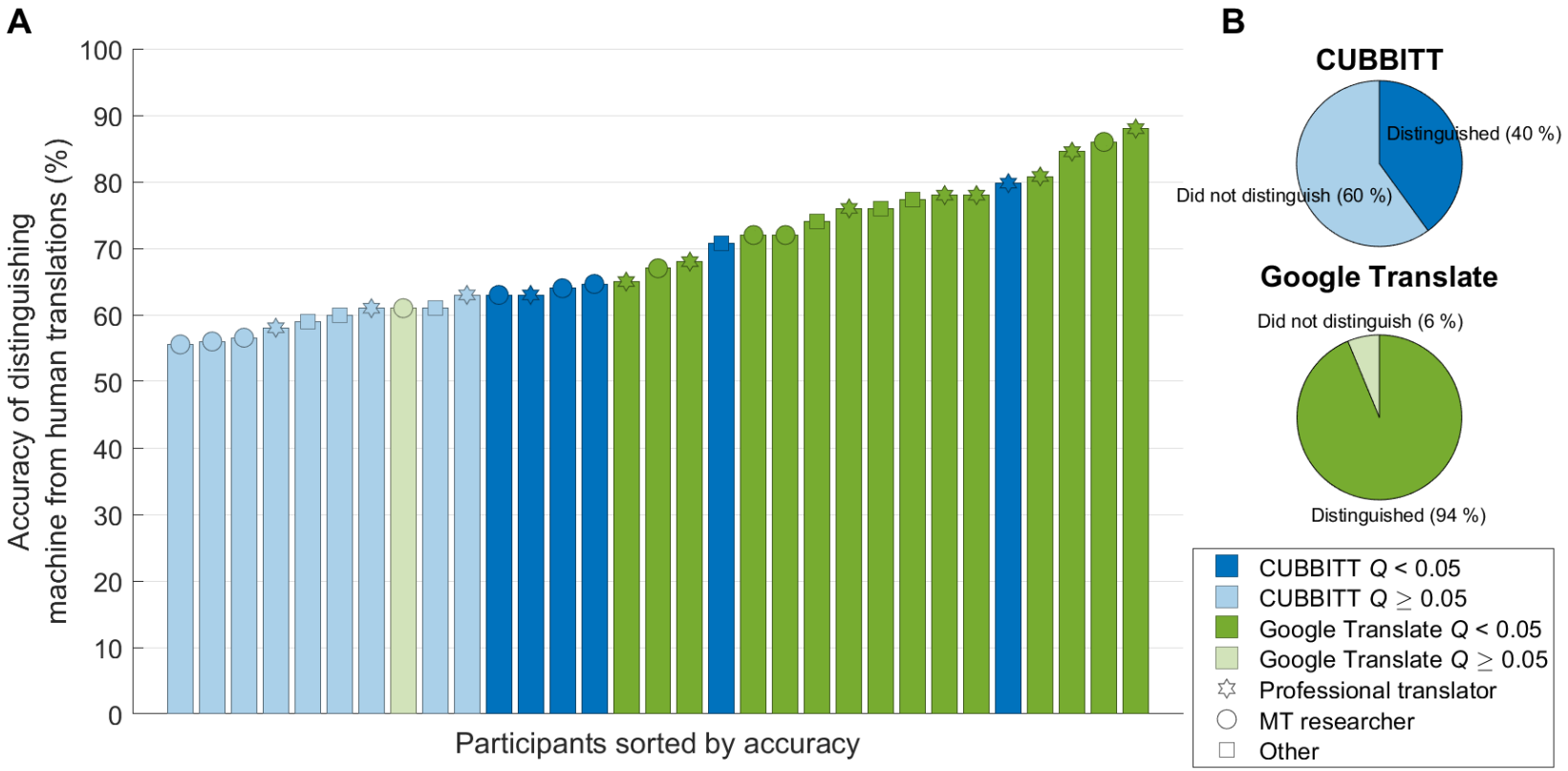
- translation of ambiguous words
- fluency (but not spelling and grammar) and
- cross-sentence context



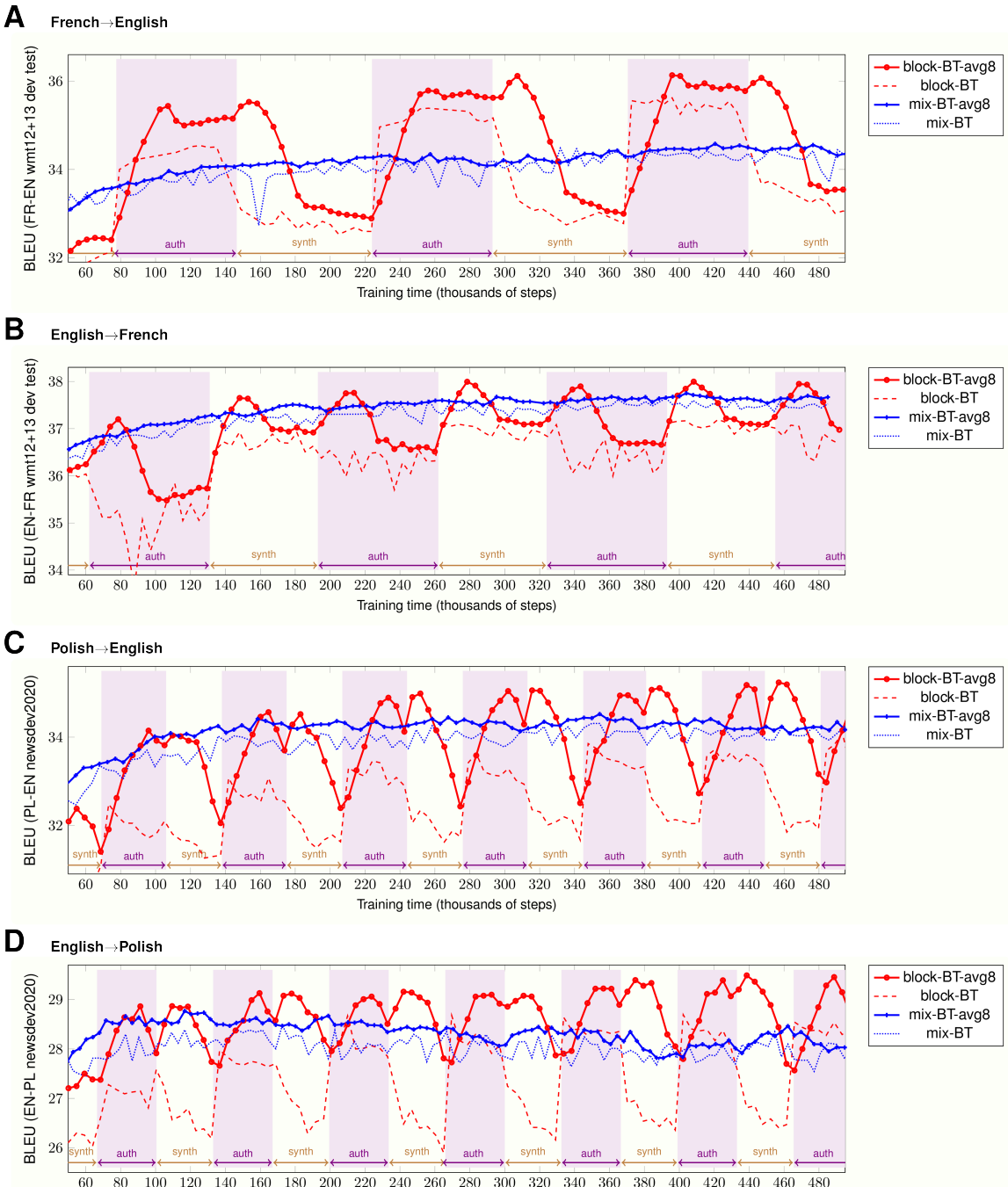
BlockBT improves the quality (+0.4 adequacy, +0.3 fluency).



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



BLEU Training curves for en-fr and en-pl



Why Block-BT works?

A

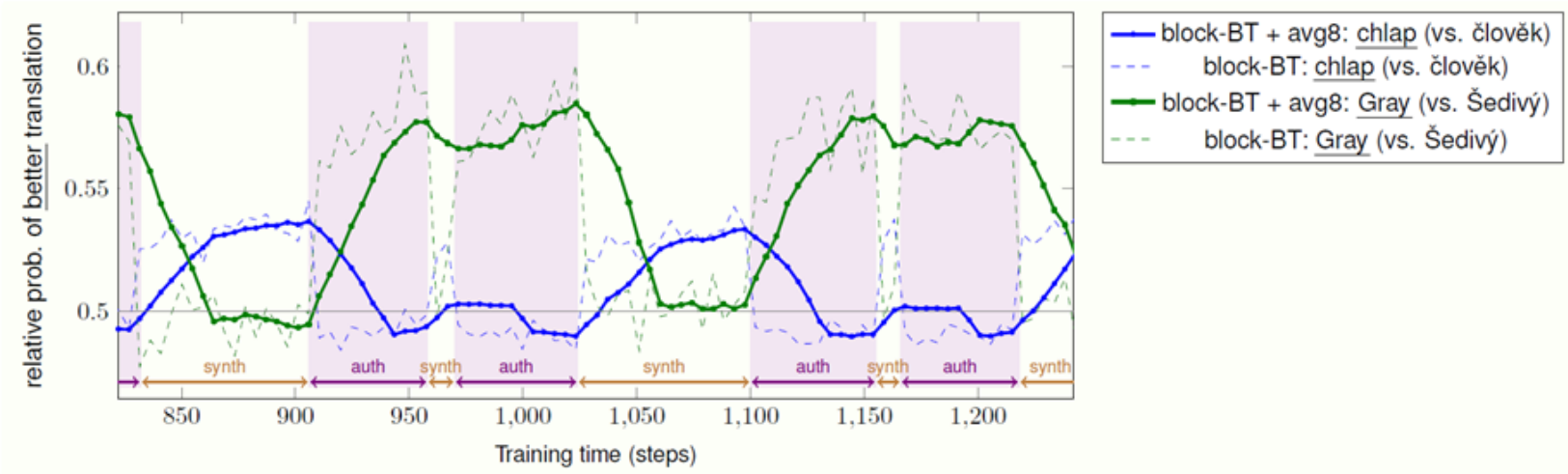
Source sentence: "He was an original **guy** and lived life to the full" said **Gray** in a statement.

SYNTH: **chlap** AUTH: **člověk** SYNTH: **Šedivý** AUTH: **Gray**

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



Translation examples

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 feťácké vejce jako činný pták.
T2018	Dobří buďte 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.
Bing	Miss o palec je Miss o míli.
Yandex	Slečna tím, že palec je vedle o míli.
Google	Chybějící palcem je míle vzdálená míle.
T2009	Slečna palec je slečna miliónu.
T2018	Slečna palce je slečna míle.
CUBBITT	Minutí o centimetr je o kilometr.

Birds of a feather flock together.
Ptáci peří stáda dohromady.
Vrána k vráně sedá.
Vrána k vráně sedá.
Ptáci v bederním hejnu spolu.
Ptáci péřového hejna spolu.
Vrána k vráně sedá.

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