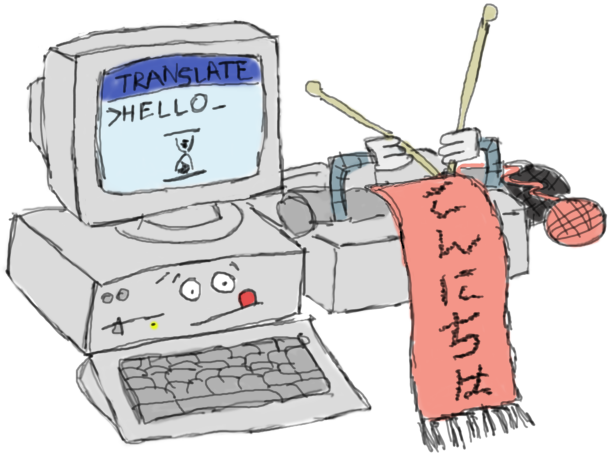
O strojovém překladu

Mgr. Martin Popel, Ph.D.
Ústav formální a aplikované lingvistiky,
Matematicko-fyzikální fakulta, Univerzita Karlova
http://ufal.cz/martin-popel



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.

Čím se zabývá počítačová lingvistika?

Kontrola ~~překlepů~~ gramatiky

Chlapci šly.

Chlapec šli do školy.

Kontrola ~~překlepů~~ gramatiky

Dívce nešly hodinky.

Chlapci šly.

Chlapec šli do školy.

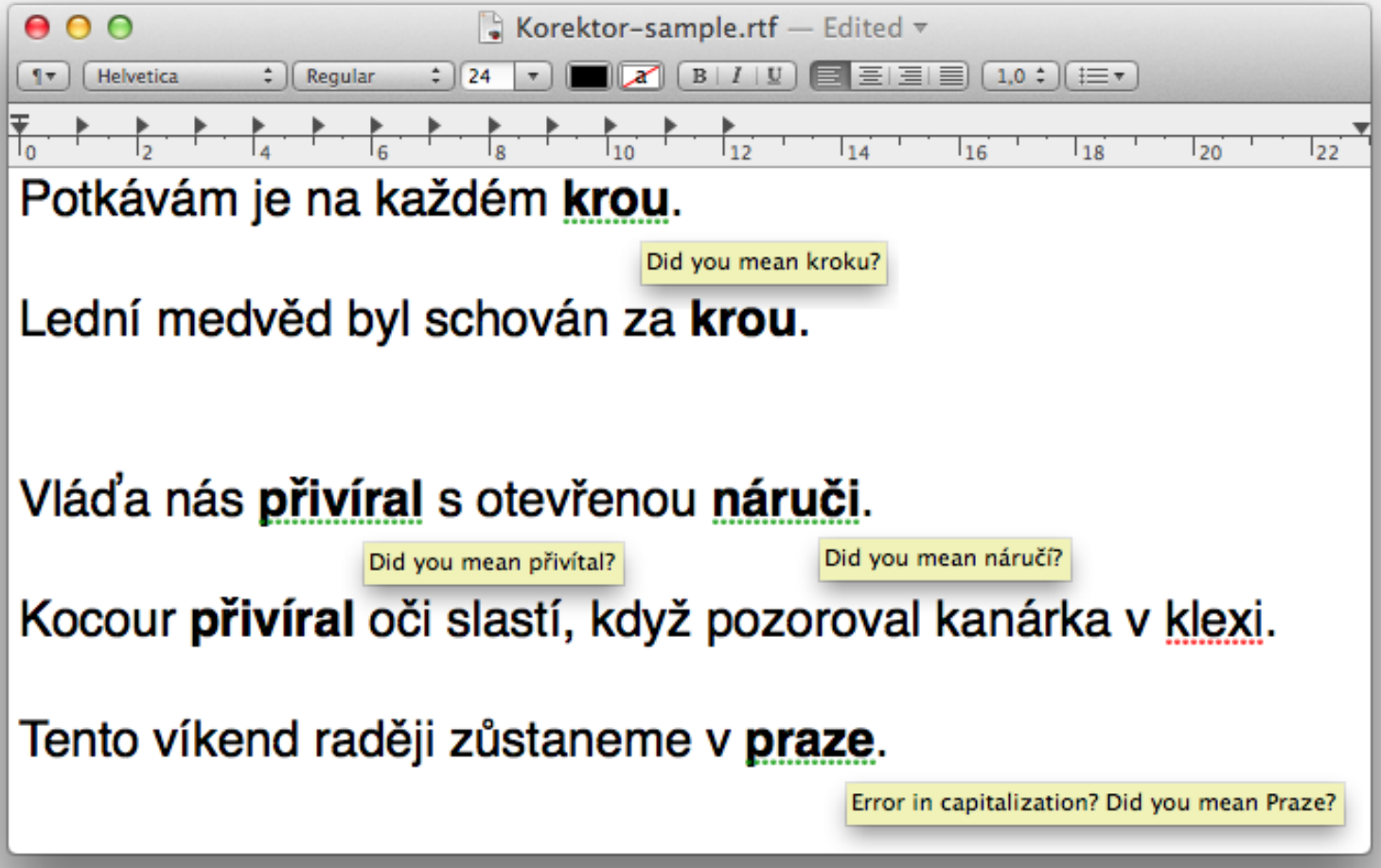
Kontrola ~~překlepů~~ gramatiky

Dívce nešly hodinky.
Kdo kam co donesl?

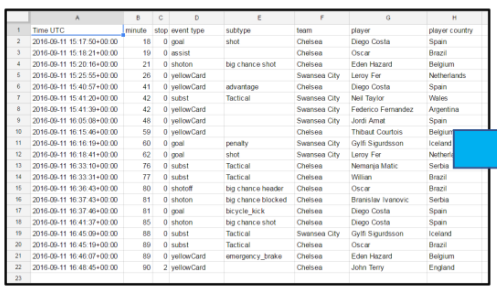
Chlapci šly.
Chlapec šli do školy.

Kontrola ~~překlepů~~ gramatiky

http://ufal.cz/korektor

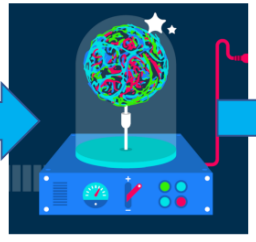


zpráv přibývá (FB, Twitter), novinářů ubývá, zisky klesají,
čtenáři chtějí zprávy personalizované a hned



	Time UTC	minute	stop	event type	subtype	team	player	player country
2	2016-08-11 15:17:50+00:00	18	0	goal	shot	Chelsea	Diego Costa	Spain
3	2016-08-11 15:18:21+00:00	19	0	assist		Chelsea	Oscar	Brazil
4	2016-08-11 15:20:16+00:00	21	0	shoton	big chance shot	Chelsea	Eden Hazard	Belgium
5	2016-08-11 15:25:55+00:00	26	0	yellowCard		Swansea City	Leroy Fer	Netherlands
6	2016-08-11 15:40:57+00:00	41	0	yellowCard	advantage	Chelsea	Diego Costa	Spain
7	2016-08-11 15:41:20+00:00	42	0	subst	Tactical	Swansea City	Neil Taylor	Wales
8	2016-08-11 15:41:39+00:00	42	0	yellowCard		Swansea City	Federico Fernandez	Argentina
9	2016-08-11 16:05:08+00:00	48	0	yellowCard		Swansea City	Jordi Amat	Spain
10	2016-08-11 16:15:46+00:00	59	0	yellowCard		Chelsea	Thibaut Courtois	Belgium
11	2016-08-11 16:16:19+00:00	60	0	goal	penalty	Swansea City	Gylfi Sigurdsson	Iceland
12	2016-08-11 16:18:41+00:00	62	0	goal	shot	Swansea City	Leroy Fer	Netherlands
13	2016-08-11 16:33:10+00:00	76	0	subst	Tactical	Chelsea	Nemanja Matic	Serbia
14	2016-08-11 16:33:31+00:00	77	0	subst	Tactical	Chelsea	Willian	Brazil
15	2016-08-11 16:36:43+00:00	80	0	shotoff	big chance header	Chelsea	Oscar	Brazil
16	2016-08-11 16:37:43+00:00	81	0	shoton	big chance blocked	Chelsea	Branislav Ivanovic	Serbia
17	2016-08-11 16:37:46+00:00	81	0	goal	bicycle_kick	Chelsea	Diego Costa	Spain
18	2016-08-11 16:41:37+00:00	85	0	shoton	big chance shot	Chelsea	Diego Costa	Spain
19	2016-08-11 16:45:09+00:00	88	0	subst	Tactical	Swansea City	Gylfi Sigurdsson	Iceland
20	2016-08-11 16:45:19+00:00	89	0	subst	Tactical	Chelsea	Oscar	Brazil
21	2016-08-11 16:46:07+00:00	89	0	yellowCard	emergency_brake	Chelsea	Eden Hazard	Belgium
22	2016-08-11 16:48:45+00:00	90	2	yellowCard		Chelsea	John Terry	England

Structured data
(real-time & historical)



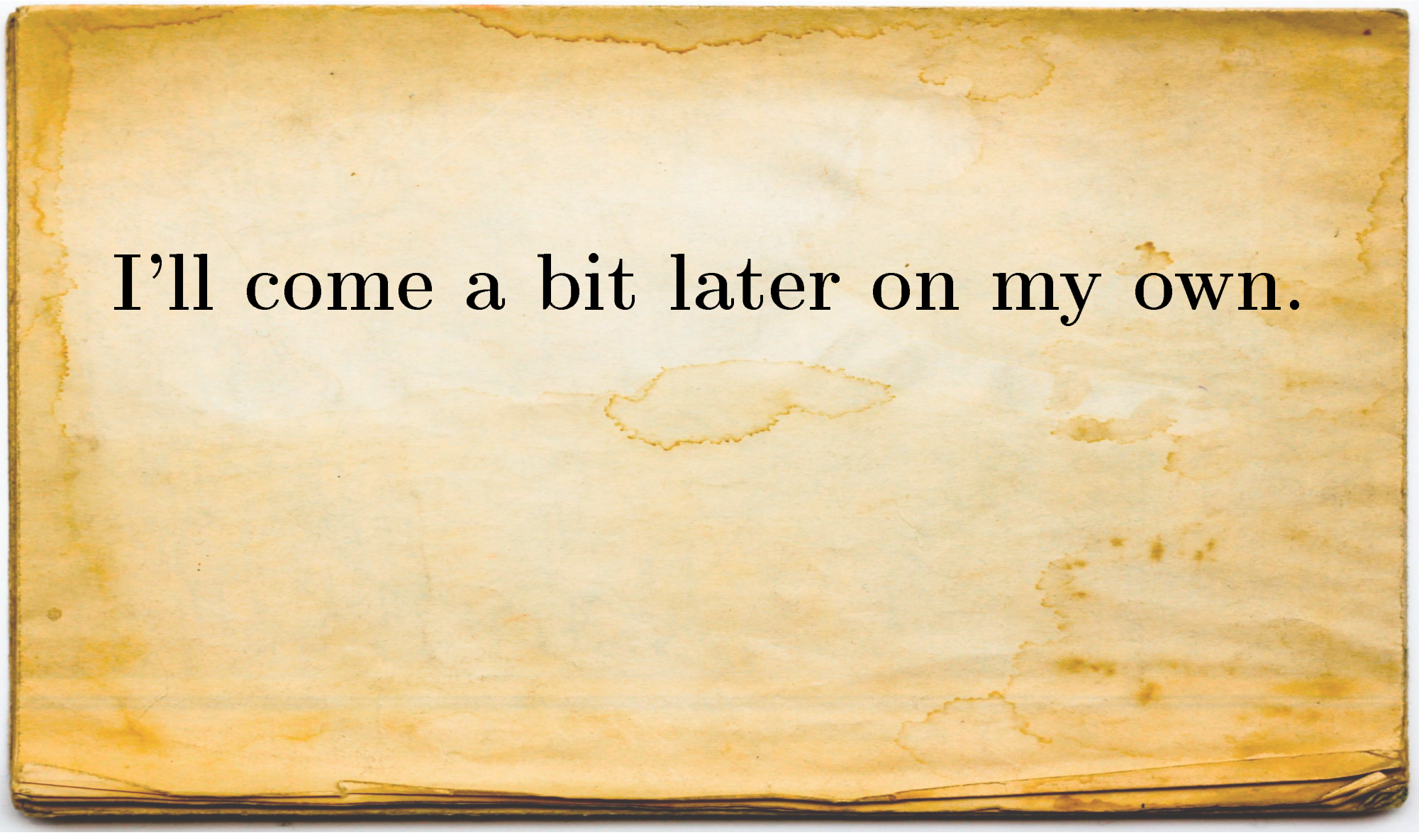
Article / Report

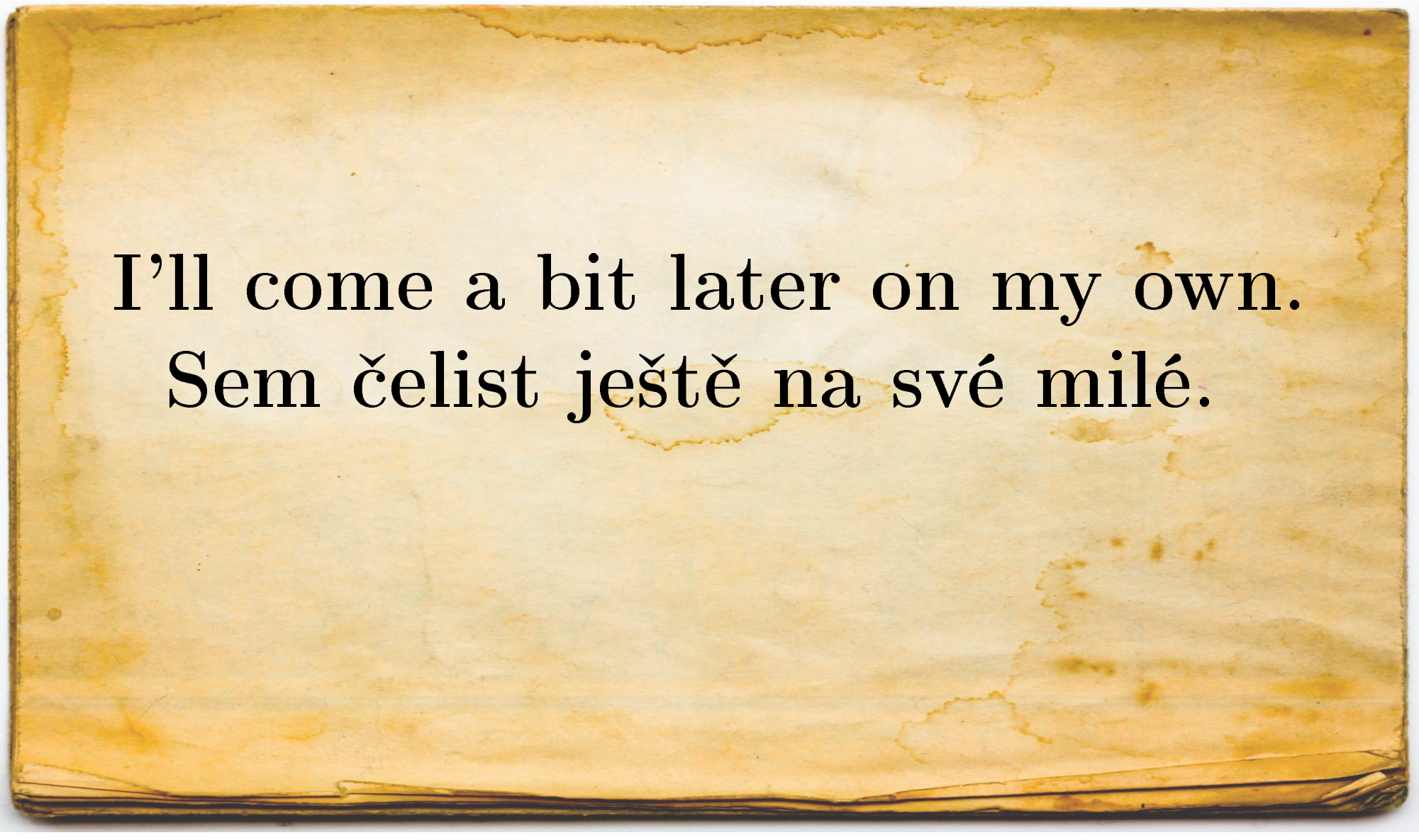
Byl by to rytíř, kde v pláně hřích vzlet,
Vědě jsem jse seheldo na přídoutně v světě si nezastavá:
„Ukryjemné, chvěla, milý nás jest

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.

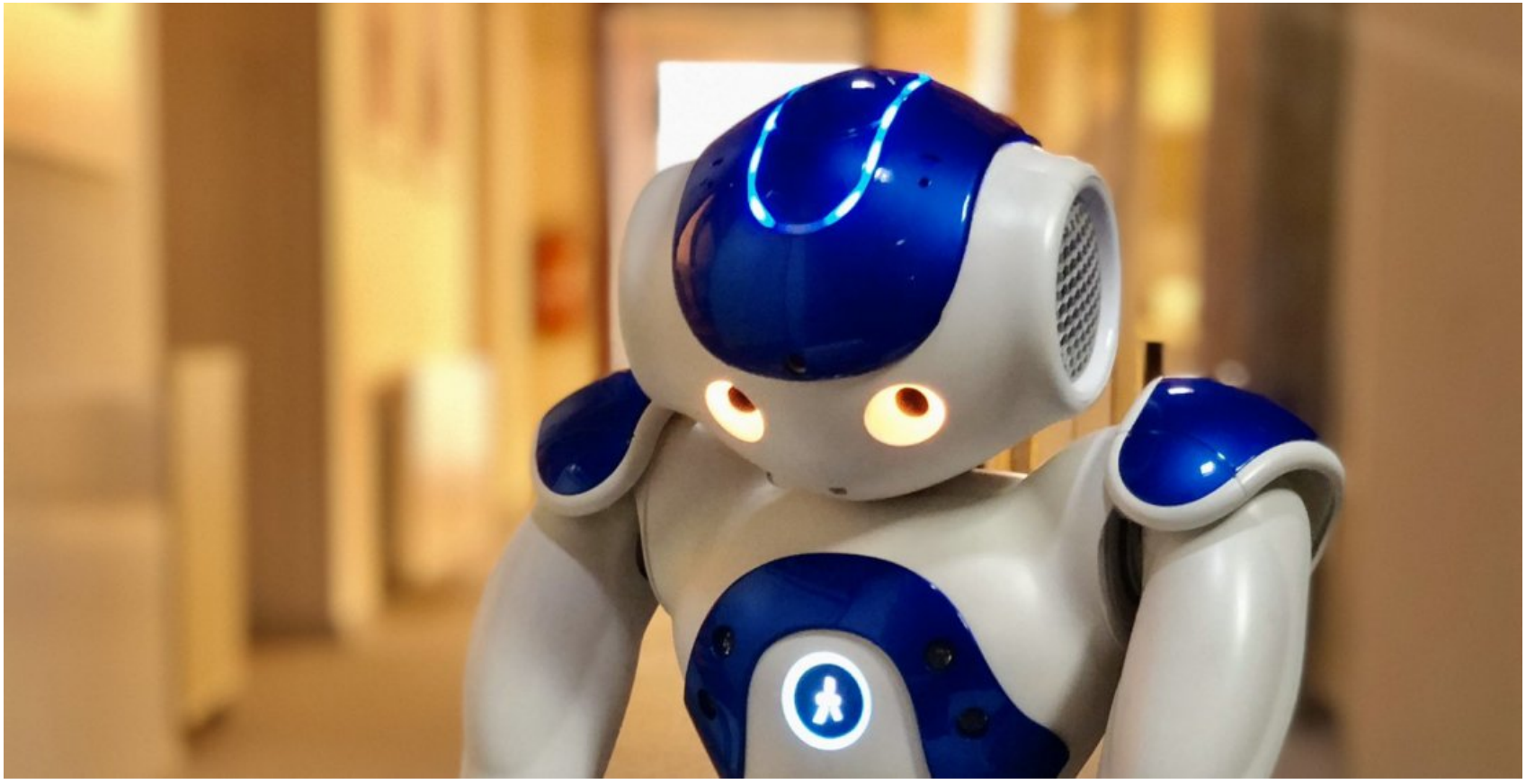

I'll come a bit later on my own.


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

Když robot píše (divadelní) hru

MFF + DAMU + Švandovo divadlo

premiéra 26. 2. 2021 (100 let po 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

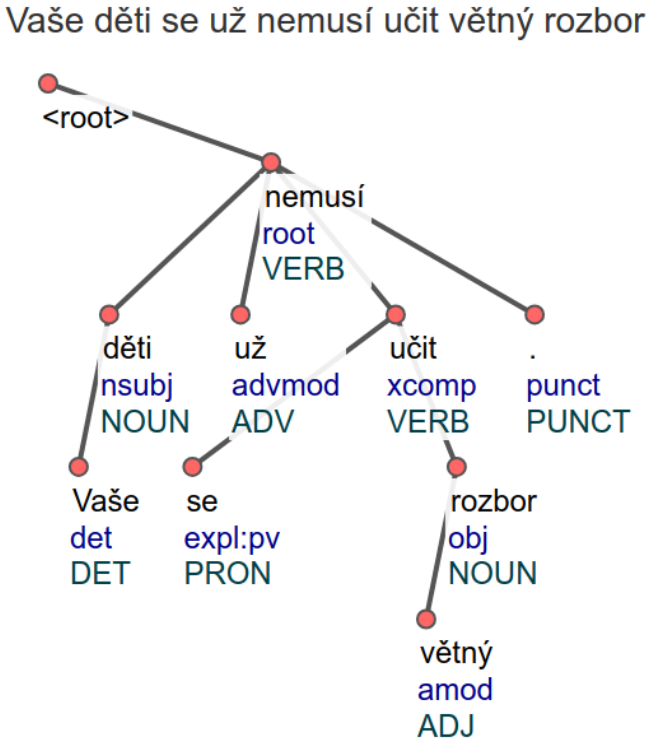
Kateřina Veselovská

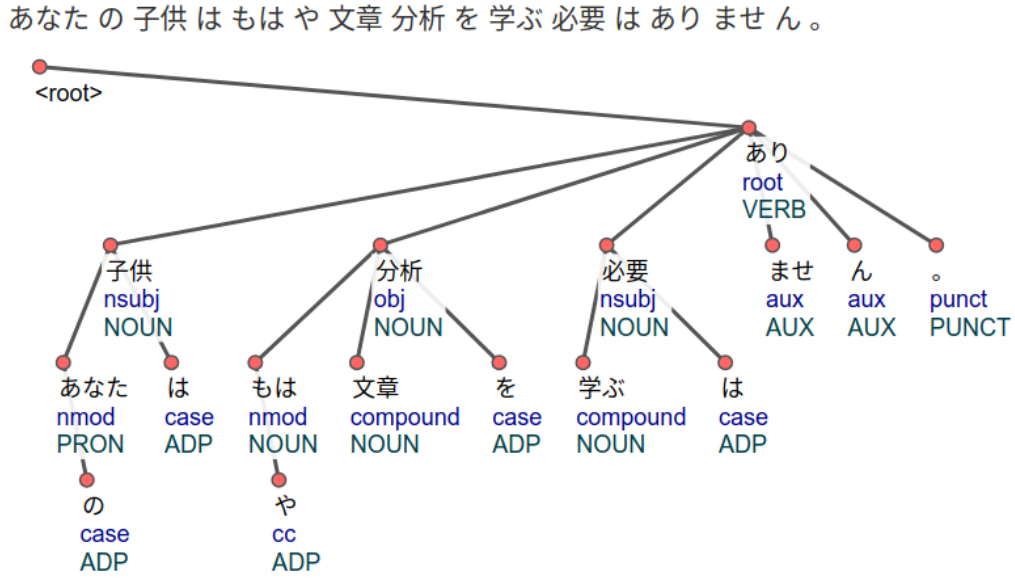
- Manažerka oddělení Data Analytics Deloitte, kde vede tým pro analýzu nestrukturovaných dat. Absolvovala doktorandský program na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze. Věnovala se vývoji softwaru pro textovou analytiku a business poradenství v projektech týkajících se oblasti nestrukturovaných 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.

větný rozbor dostupný pro 50 jazyků
přesnost pro češtinu asi 90% (85% včetně morfologie)





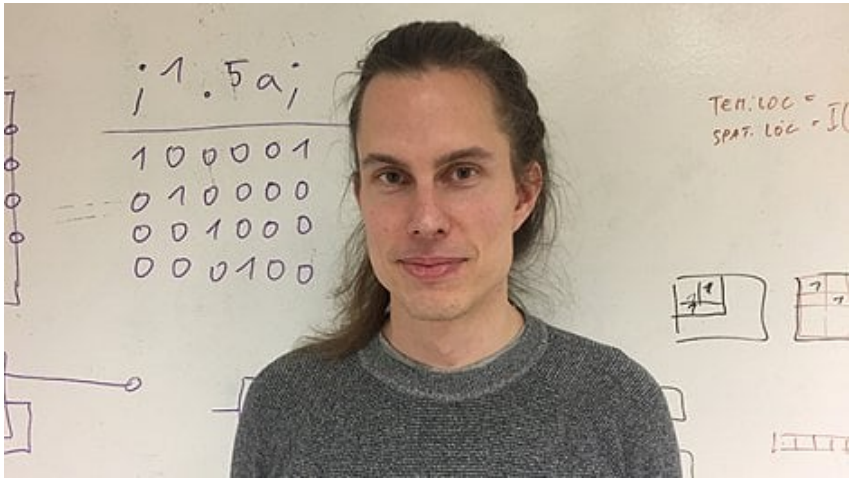
https://lindat.cz/services/udpipe/

Umíte sčítat a odčítat čísla?
A co slova a obrázky?

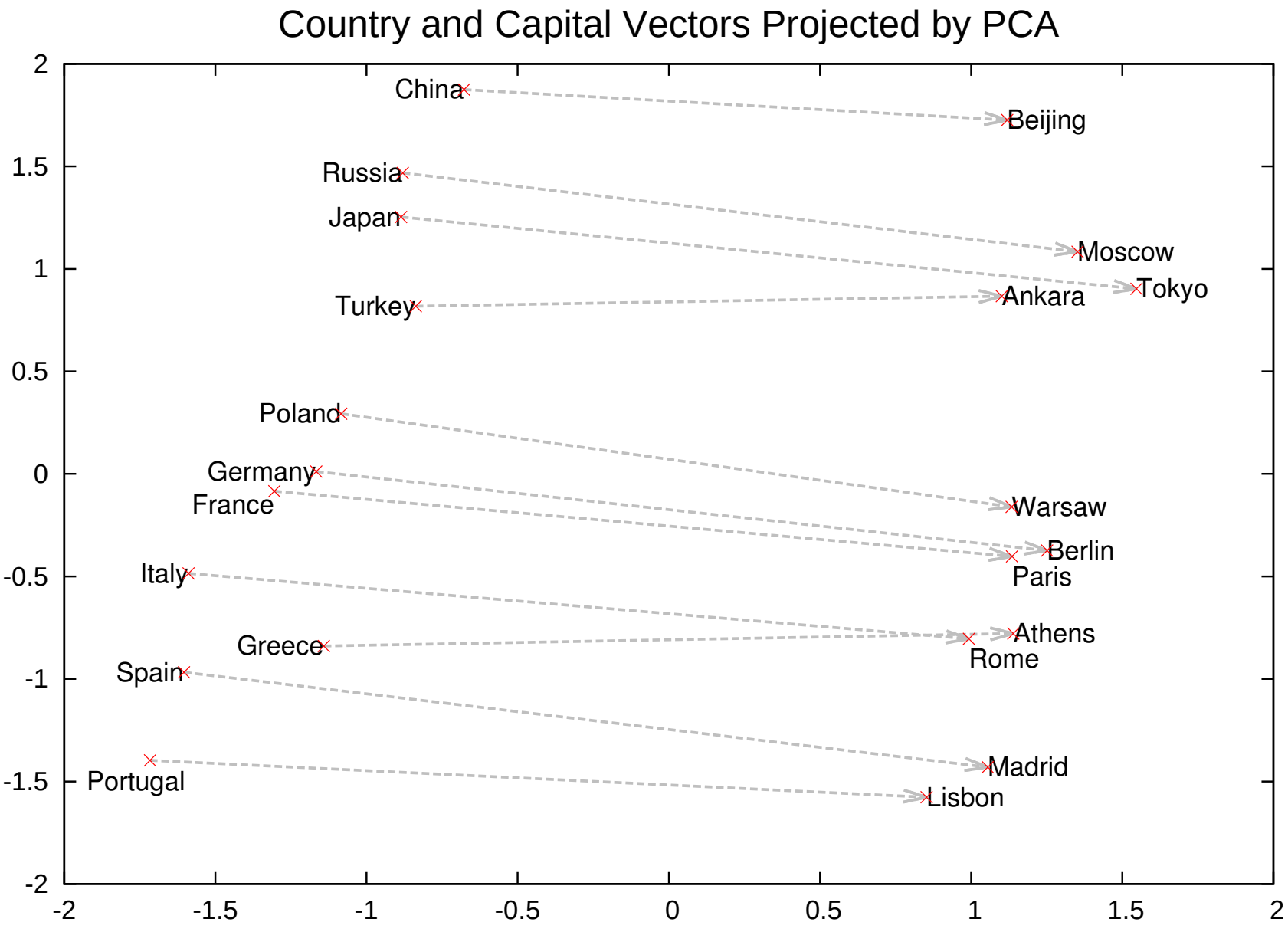
král - muž + žena = ?

král - muž + žena = královna

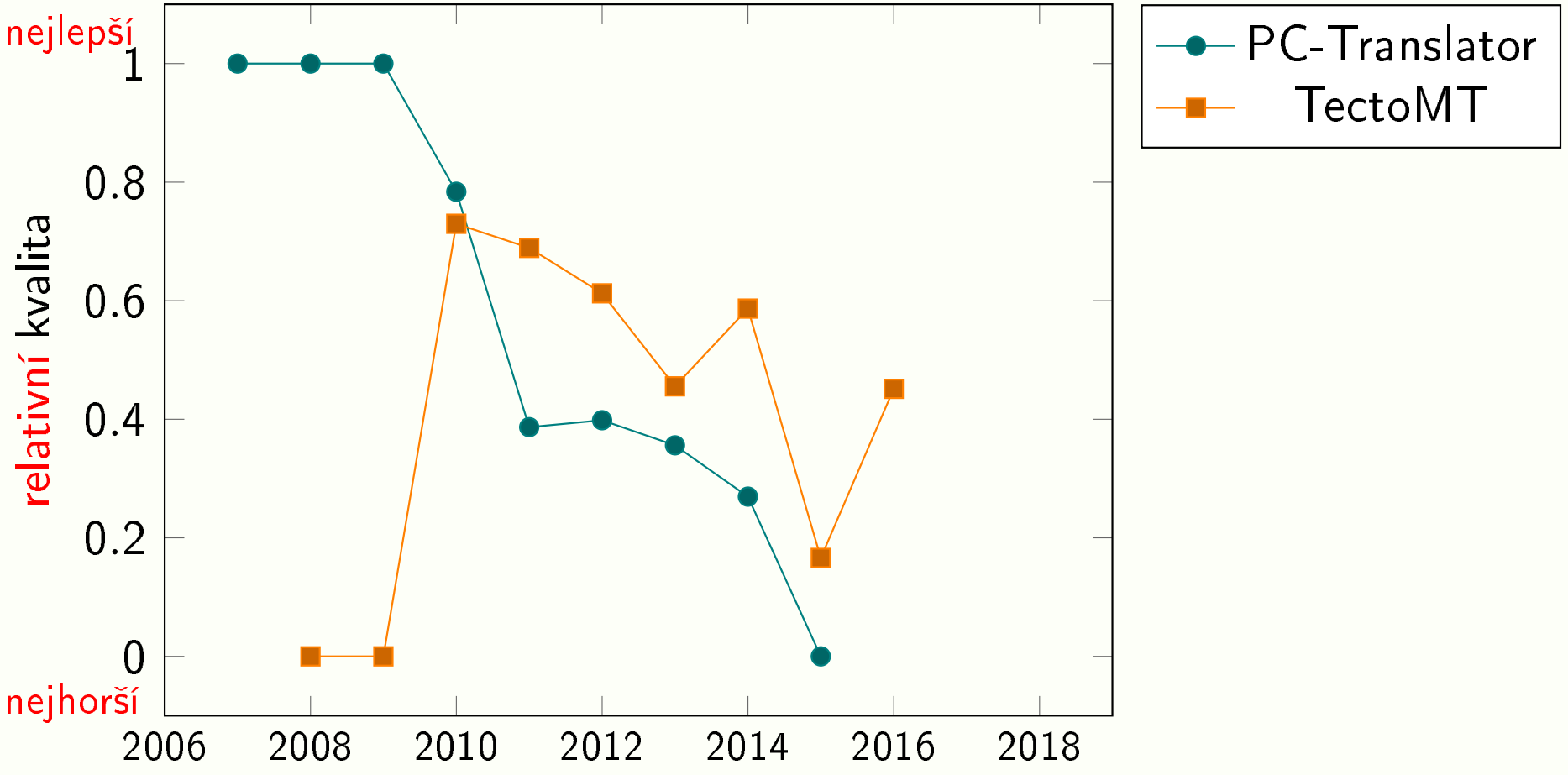
Tomáš Mikolov, 2012, word2vec



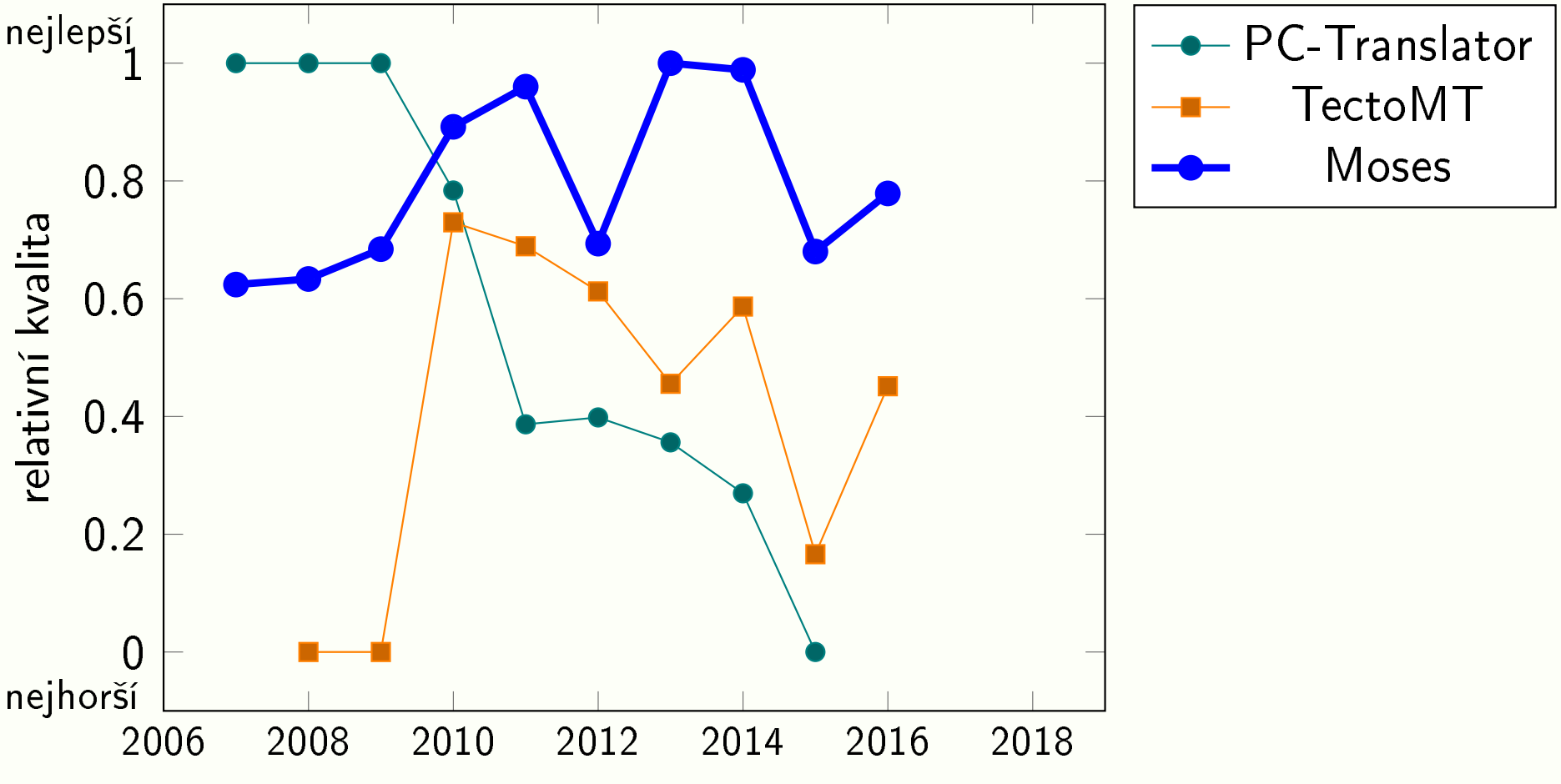
https://projector.tensorflow.org/


Country and Capital Vectors Projected by PCA
China
Beijing
Russia
Moscow
Japan
Tokyo
Turkey
Ankara
Poland
Warsaw
Germany
Berlin
France
Paris
Italy
Rome
Athens
Greece
Spain
Madrid
Portugal
Lisbon
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

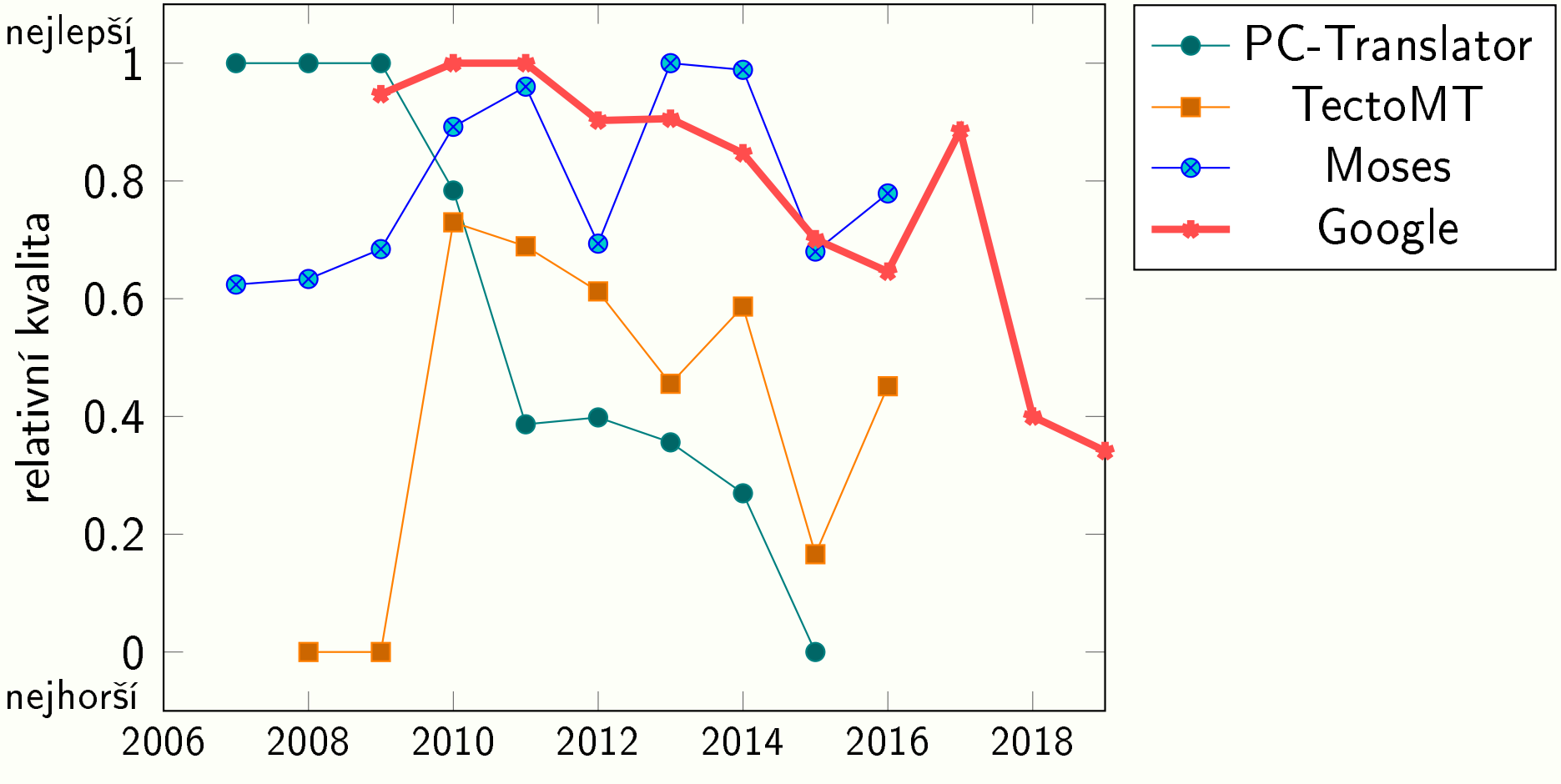
Strojový překlad angličtina→čeština 2007–2019



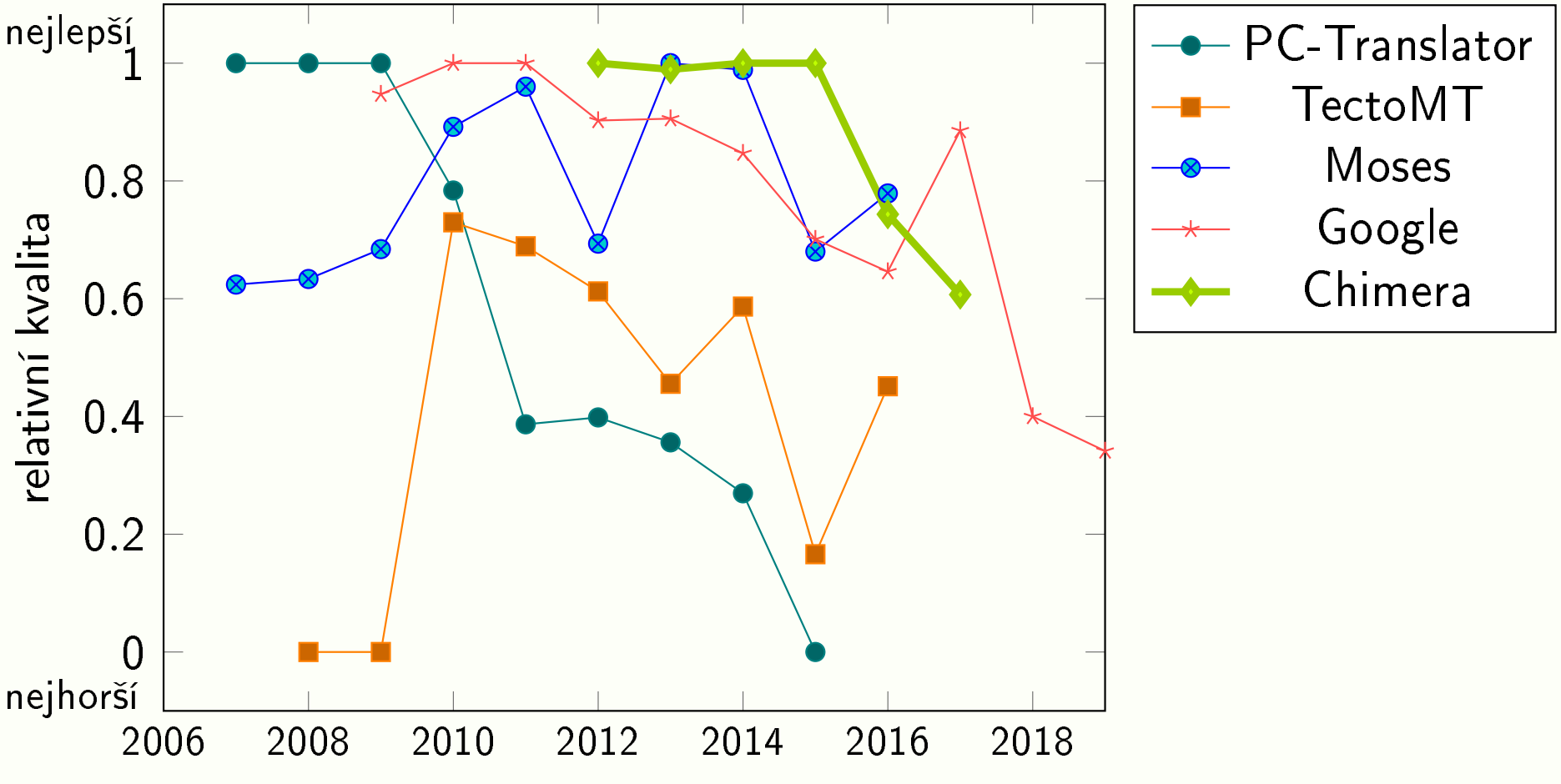
Strojový překlad angličtina→čeština 2007–2019



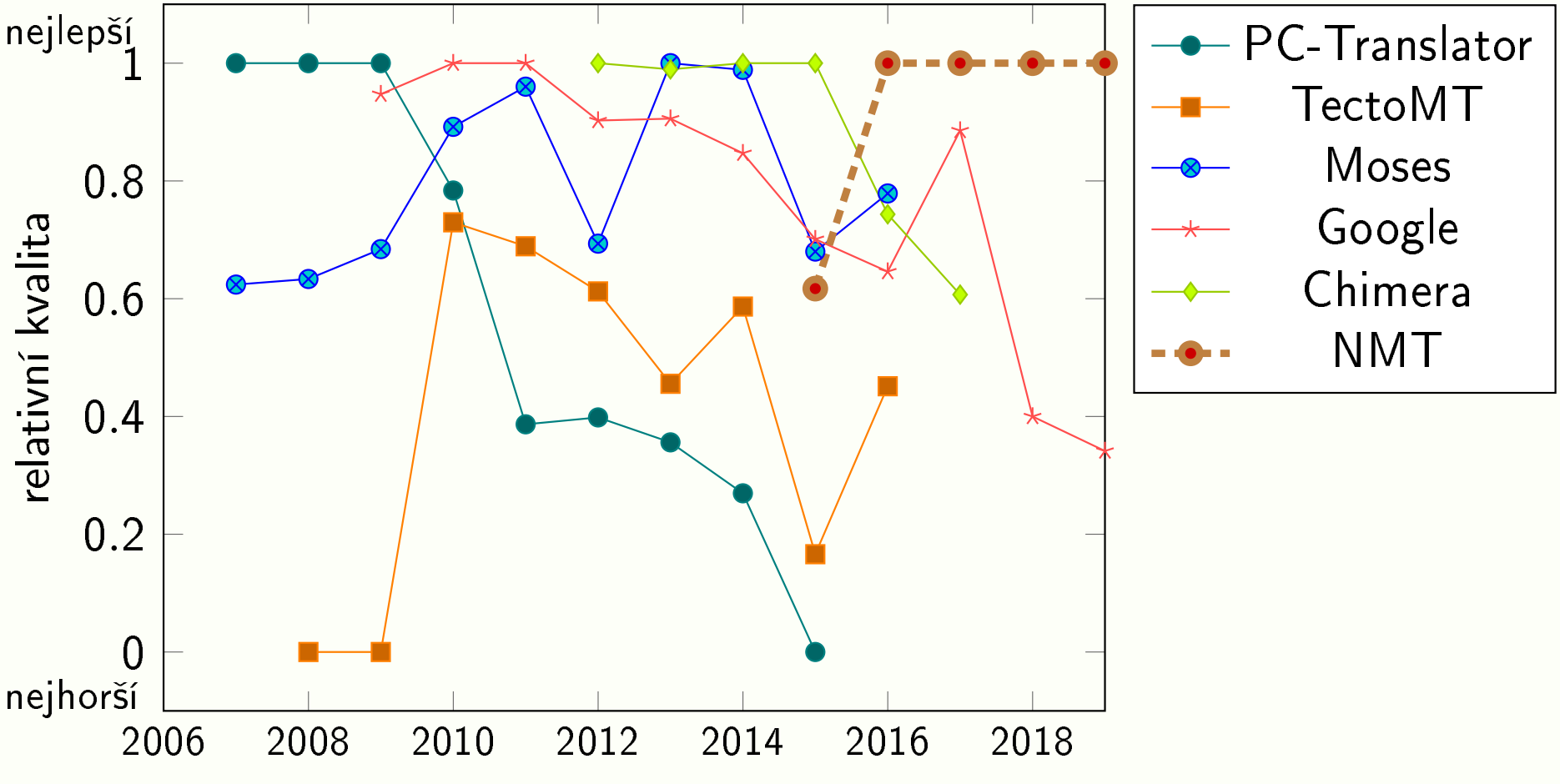
Strojový překlad angličtina→čeština 2007–2019



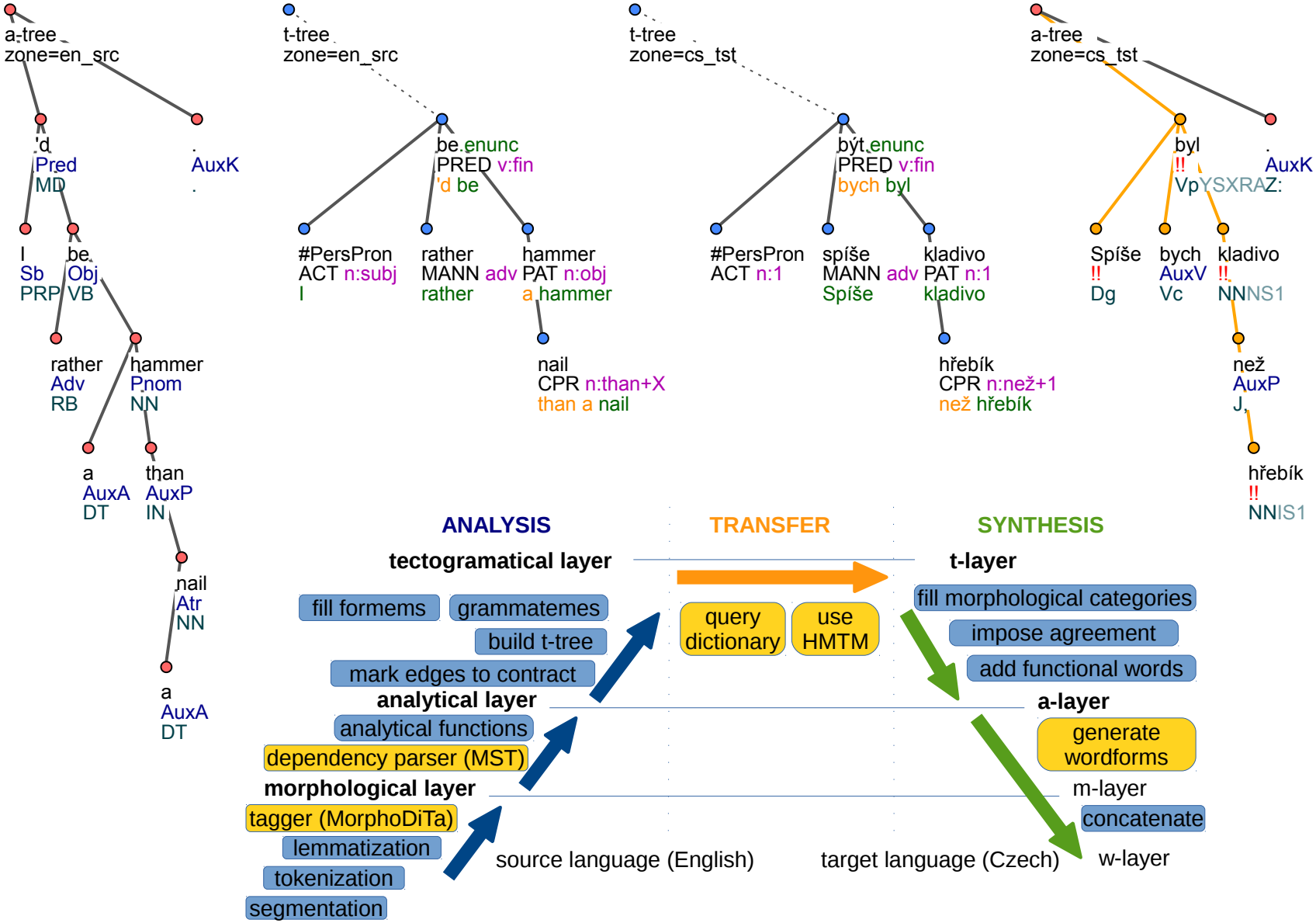
Strojový překlad angličtina→čeština 2007–2019



Strojový překlad angličtina→čeština 2007–2019



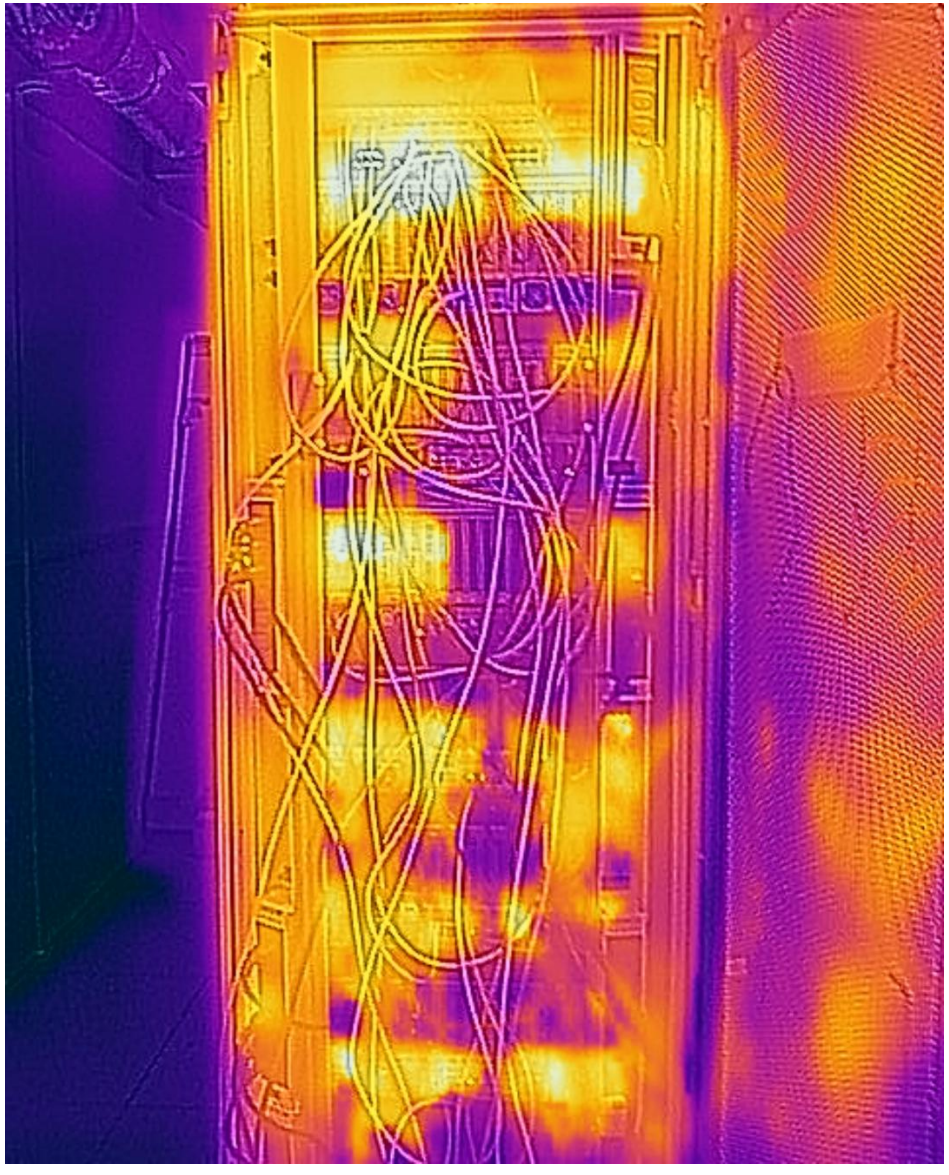
Hloubkově-syntaktický překladač TectoMT



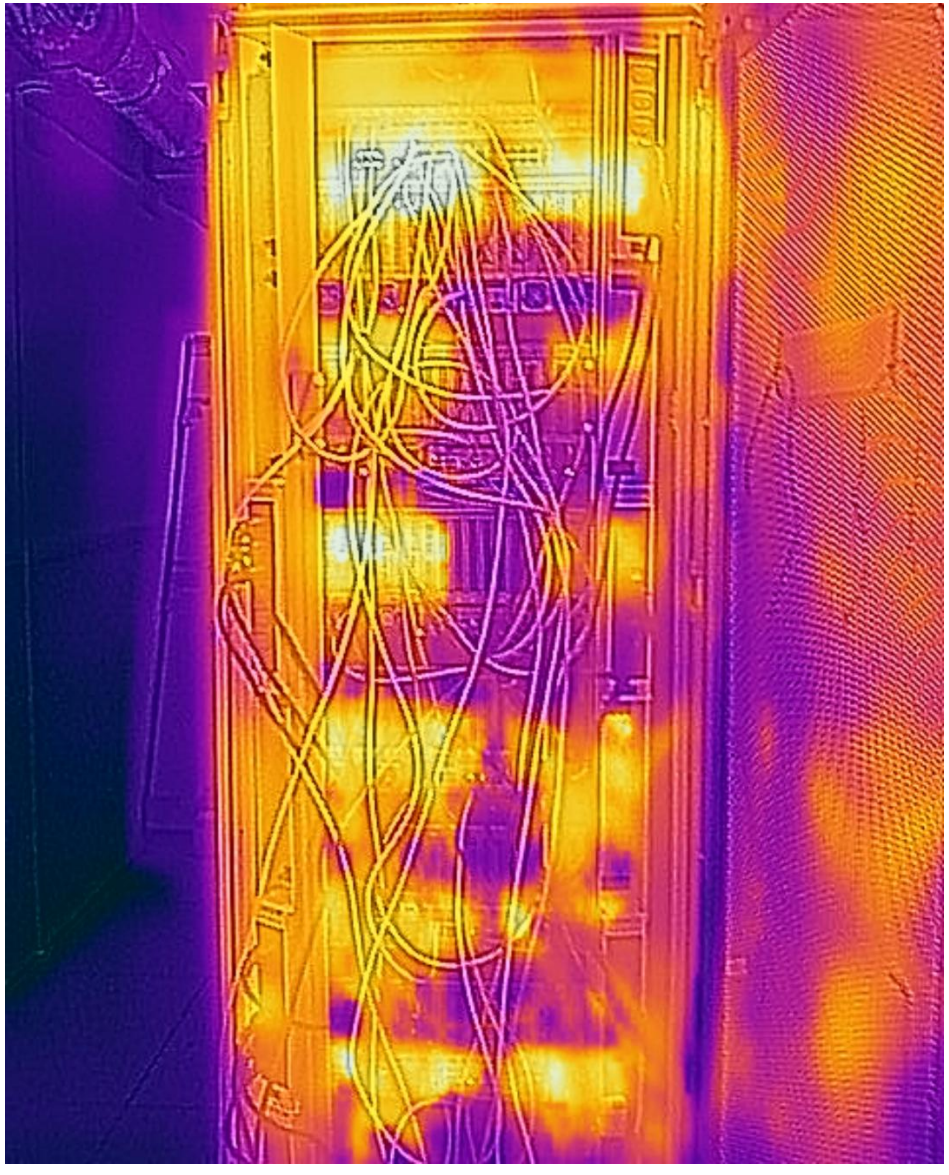
I'd rather be a hammer than a nail.

Spíše bych byl kladivo než hřebík/nehet.

Co je to?



40 GPU (GeForce GTX 1080 Ti, 12 miliard tranzistorů)

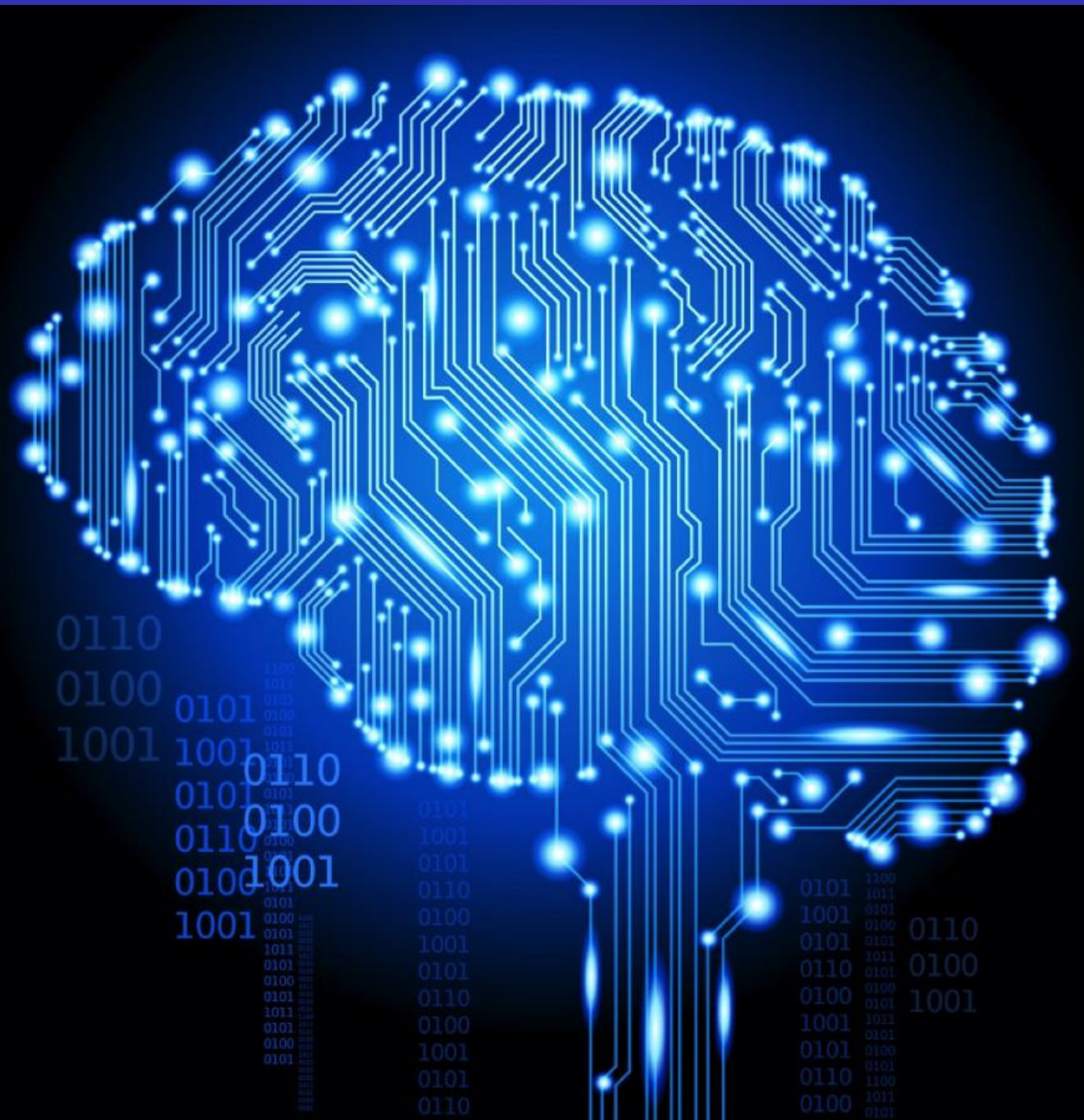


Umělá inteligence a její podobory/vývoj

umělá inteligence
~1950

strojové učení
~1980

hluboké učení
~2010

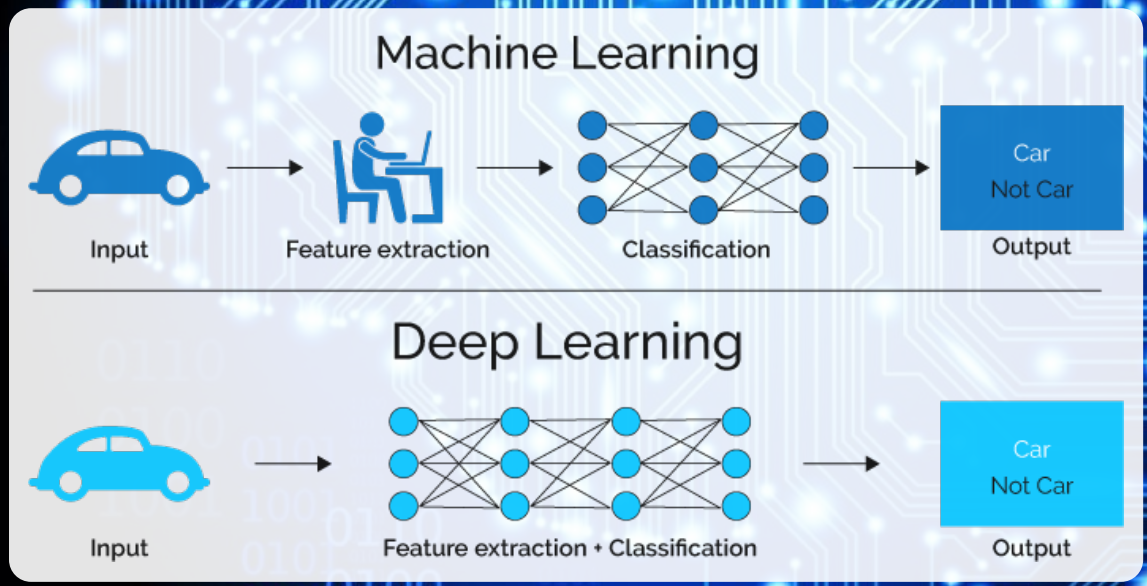


Umělá inteligence a její podobory/vývoj

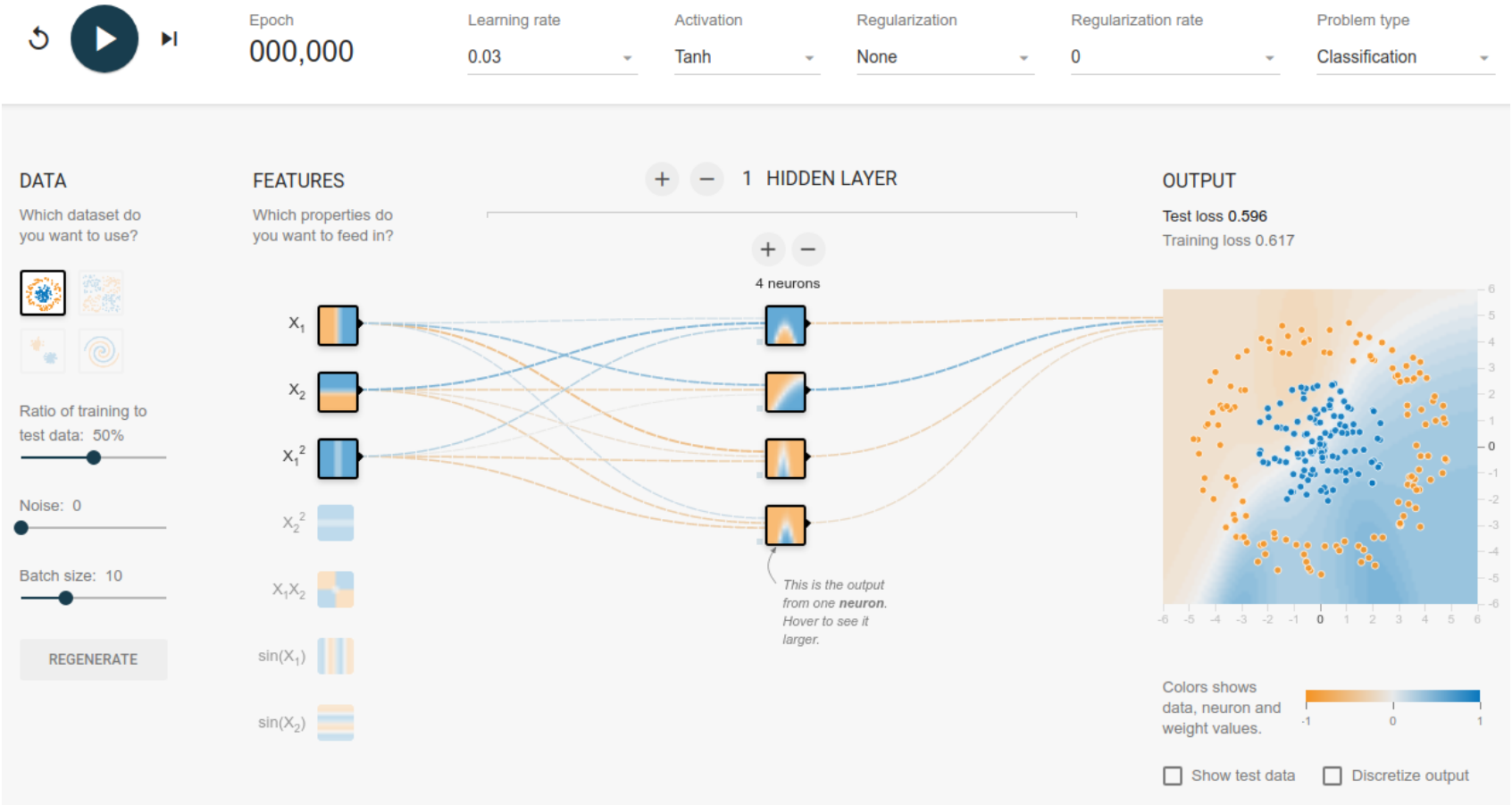
umělá inteligence
∼1950

strojové učení
∼1980

hluboké učení
∼2010

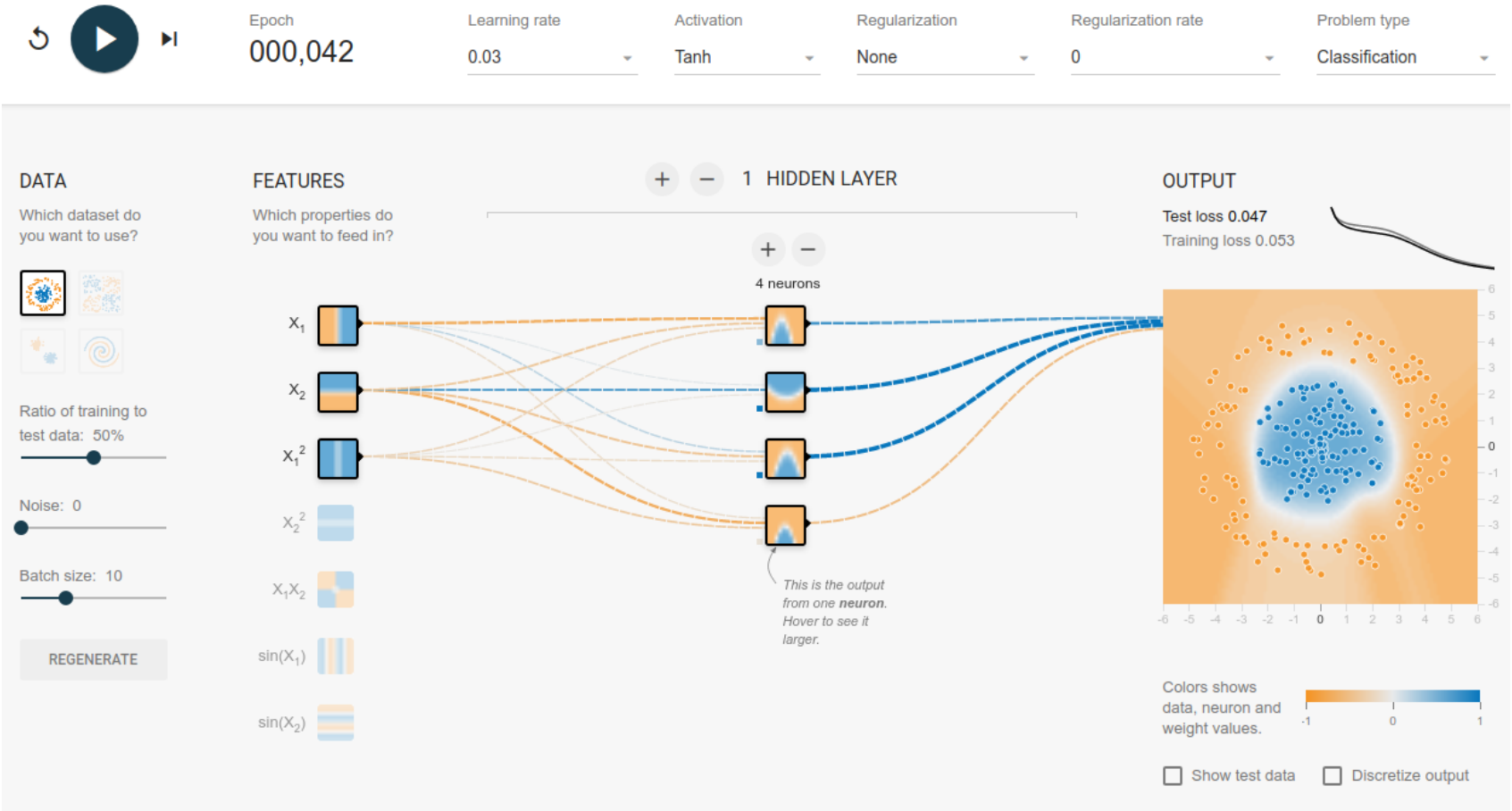


jednoduchá architektura (16 parametrů)



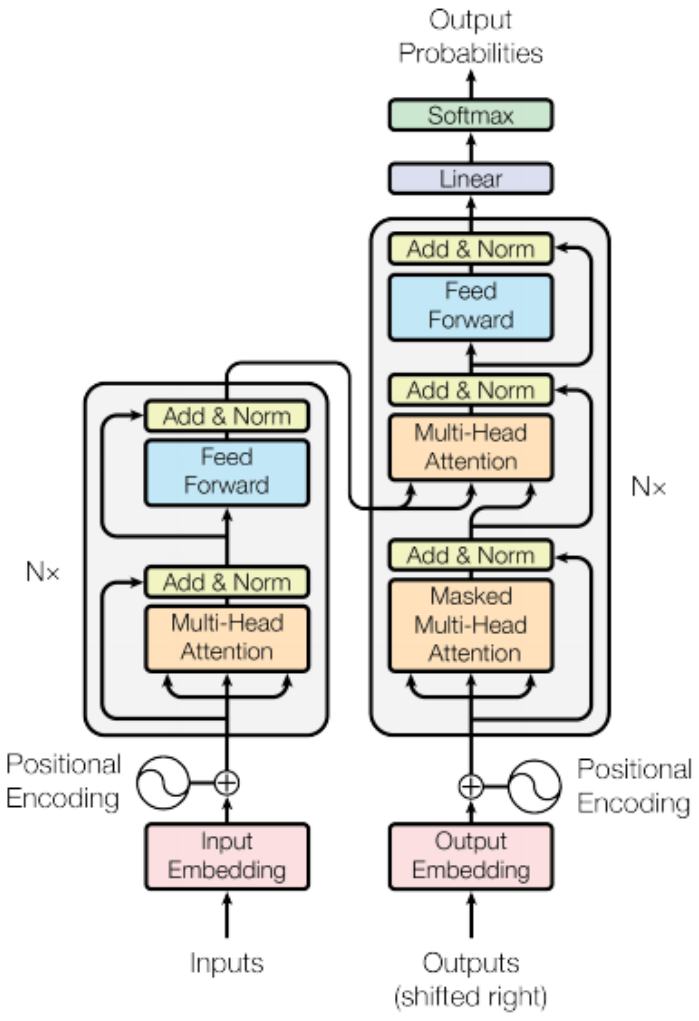
https://playground.tensorflow.org

jednoduchá architektura (16 parametrů)

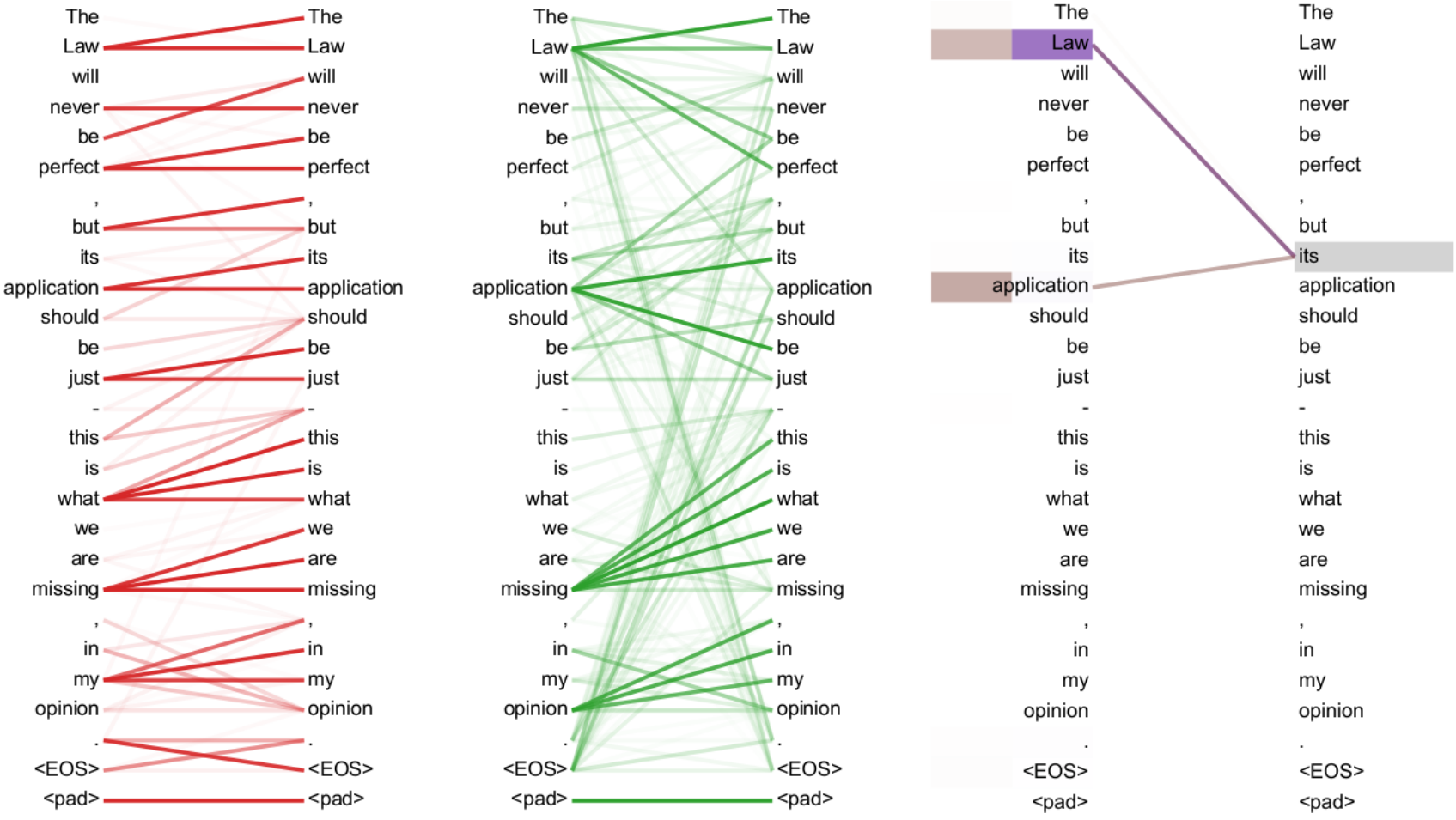


https://playground.tensorflow.org

architektura Transformer (213 milionů parametrů)

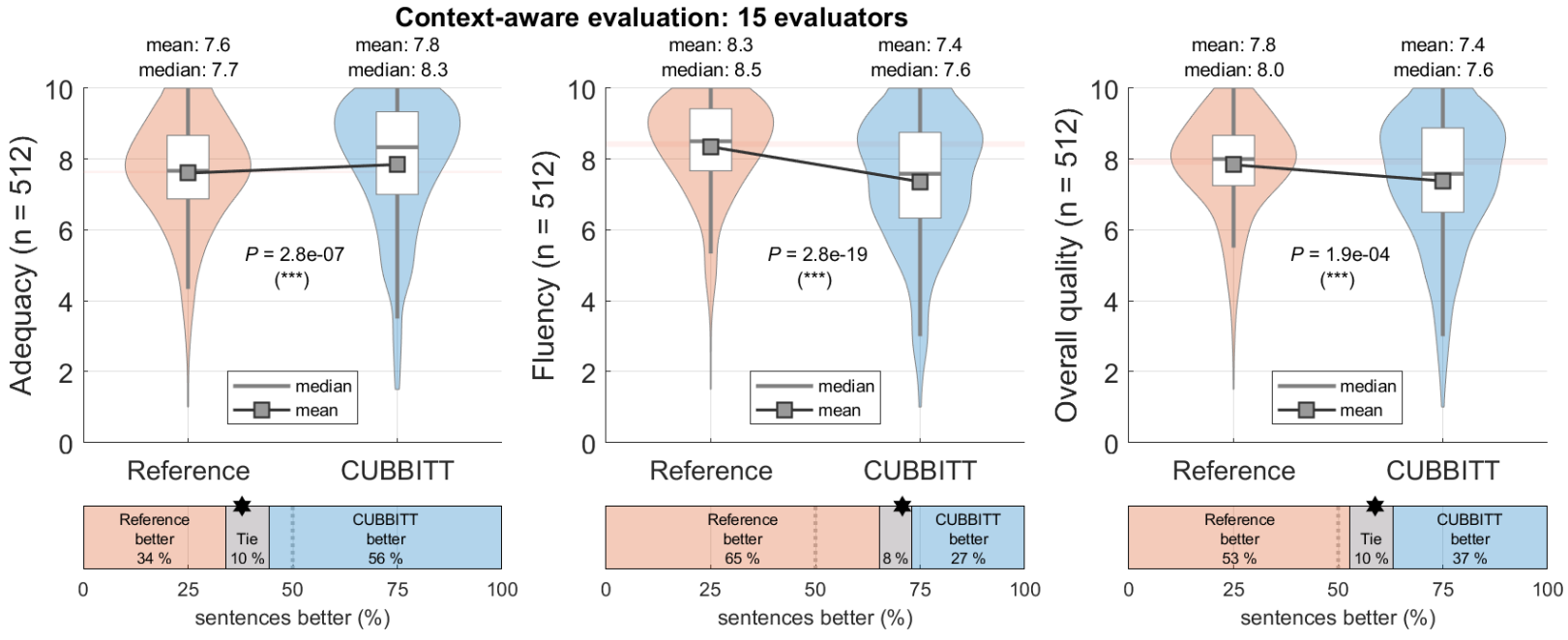


síť se sama učí důležité vztahy mezi slovy (díky self-attention)



Vyhodnocení anglicko-českého překladu

56 % vět přeložil přesněji náš překladač CUBBITT
34 % profesionální překladatelská agentura



Zkuste si CUBBITT sami:
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

6273 Accesses | **76** Altmetric | Metrics

Rizika strojového překladu

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)

Rizika strojového překladu

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)

denně přes 100 miliard slov (Google, Microsoft, Baidu, Amazon,...)

velikost trhu za rok 2020: 650 milionů dolarů

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.
DeepL	Jana is a woman. **She** works as a conductor.

Source	Nikdo nedokáže pochopit, proč Pavla dělá to, co dělá. Pracuje jako průvodčí.
CUBBITT-doc	No one can understand why Pavla does what she does. She works as a conductor.
DeepL	No one can understand why Paul does what **he** does. **She** works as a conductor.

Ukázky překladu

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.

Zkuste si CUBBITT sami: https://lindat.cz/cubbitt