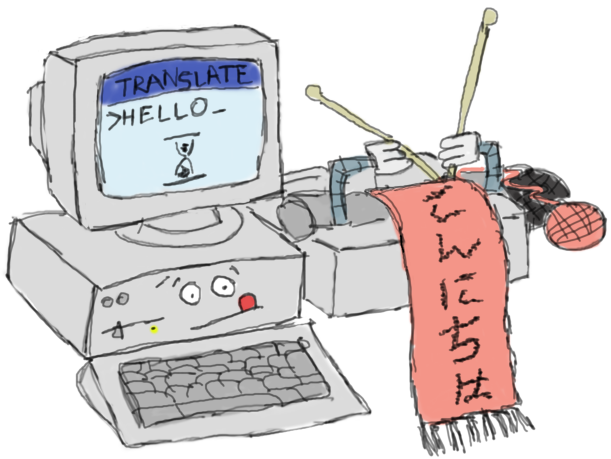
Strojový překlad a umělá inteligence

Mgr. Martin Popel, Ph.D.
Ústav formální a aplikované lingvistiky,
Matematicko-fyzikální fakulta, Univerzita Karlova



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.
Transformer	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.

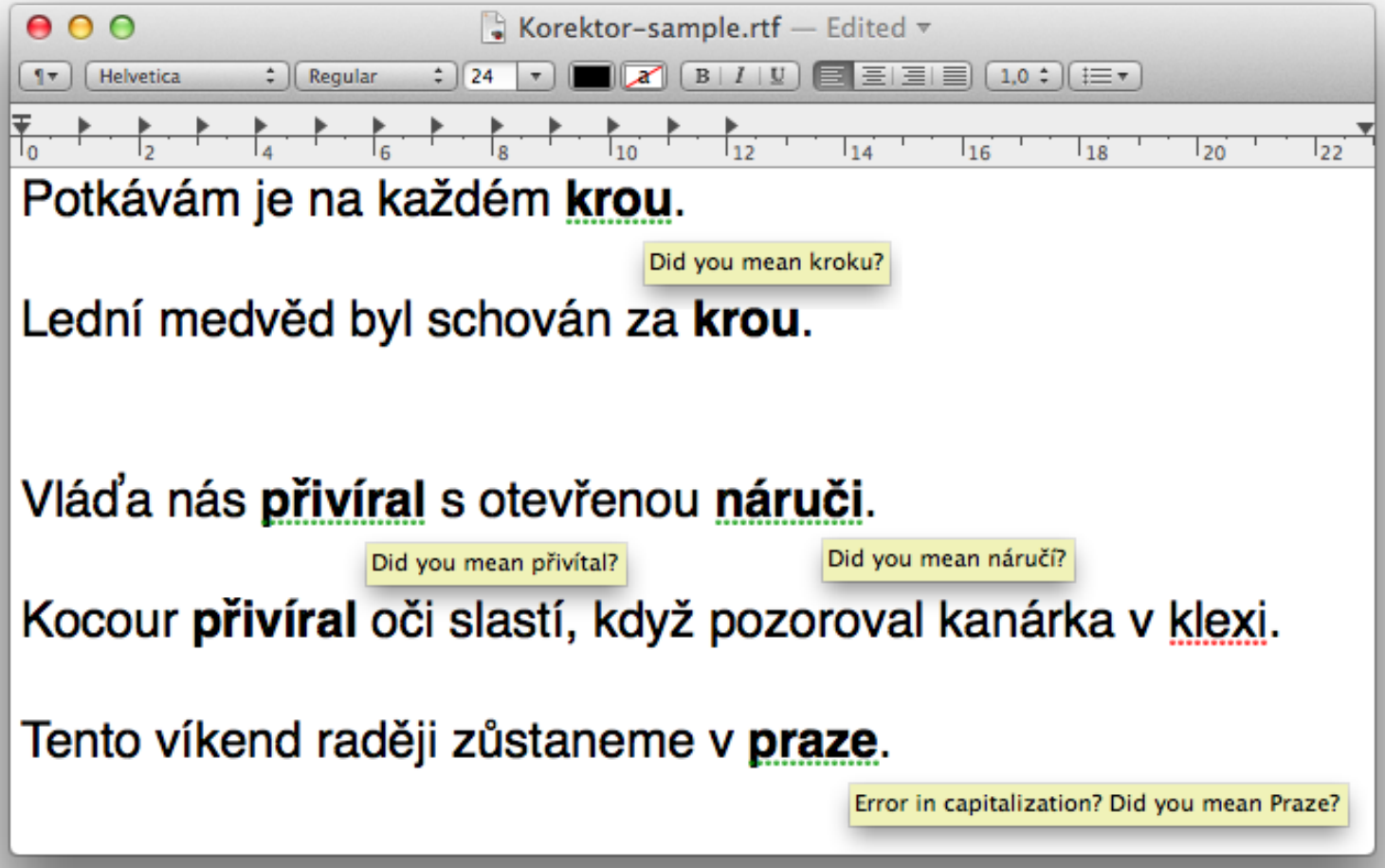
Kontrola ~~překlepů~~ gramatiky

Dívce nešly hodinky.
Kdo kam co donesl?

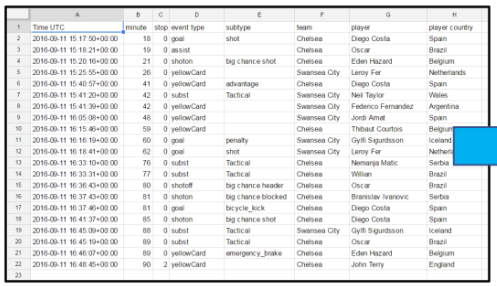
Chlapci šly.
Chlapec šli do školy.

Kontrola ~~překlepů~~ gramatiky

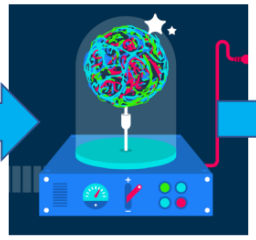
http://ufal.cz/korektor

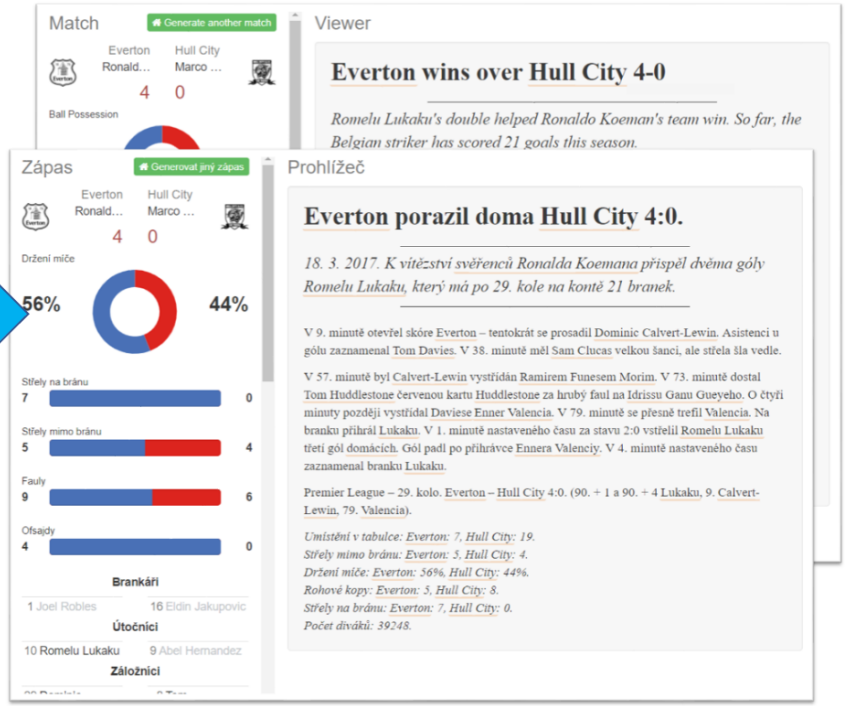


zpráv přibývá (FB, Twitter), novinářů ubývá, zisky klesají,
čtenáři chtějí zprávy personalizované a hned



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.

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

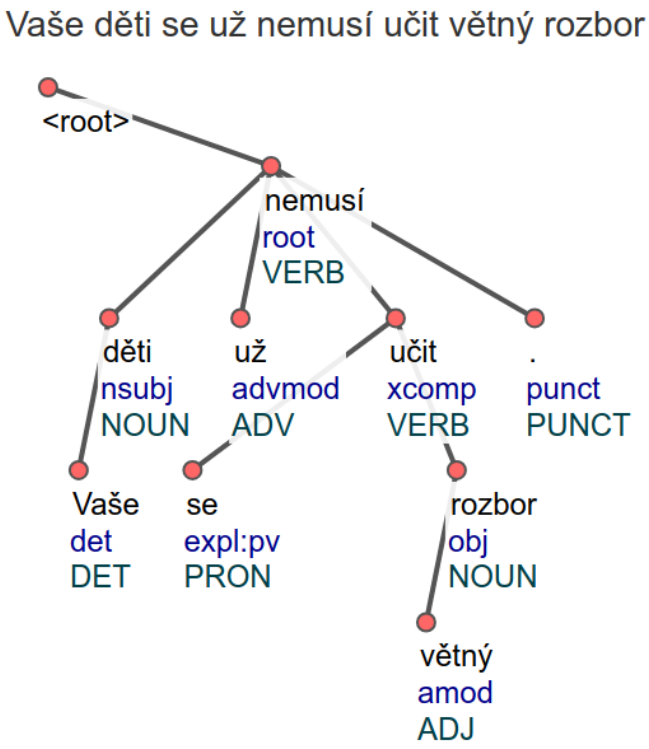
Kateřina Veselovská

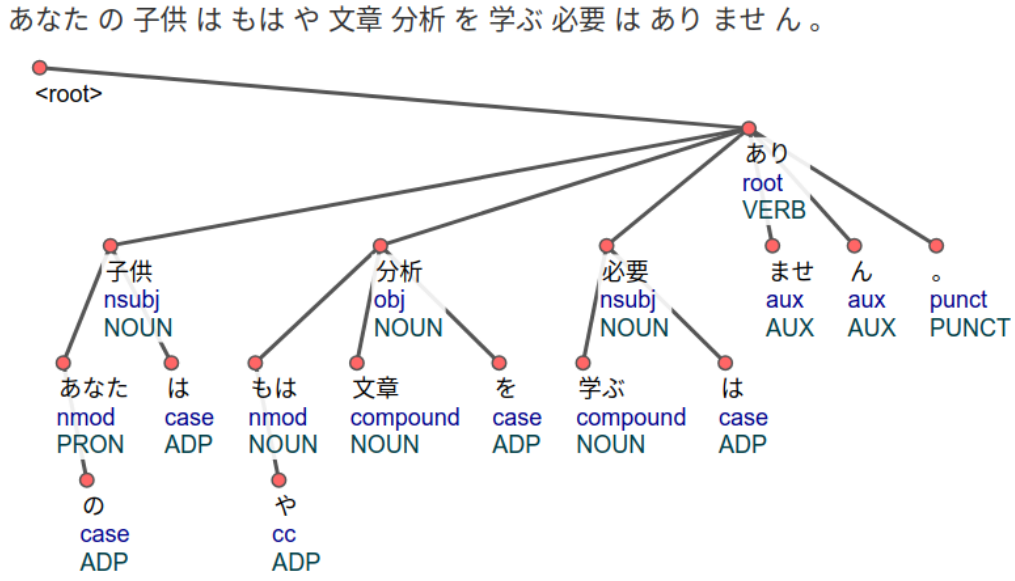
- Manažerka oddělení Data Analytics Deloitte, kde vede tým pro analýzu nestrukturovaných dat. Absolvovala doktorandský program na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze. Věnovala se vývoji softwaru pro textovou analytiku a business poradenství v projektech týkajících se oblasti nestrukturovaných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





http://lindat.cz/services/udpipe/

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

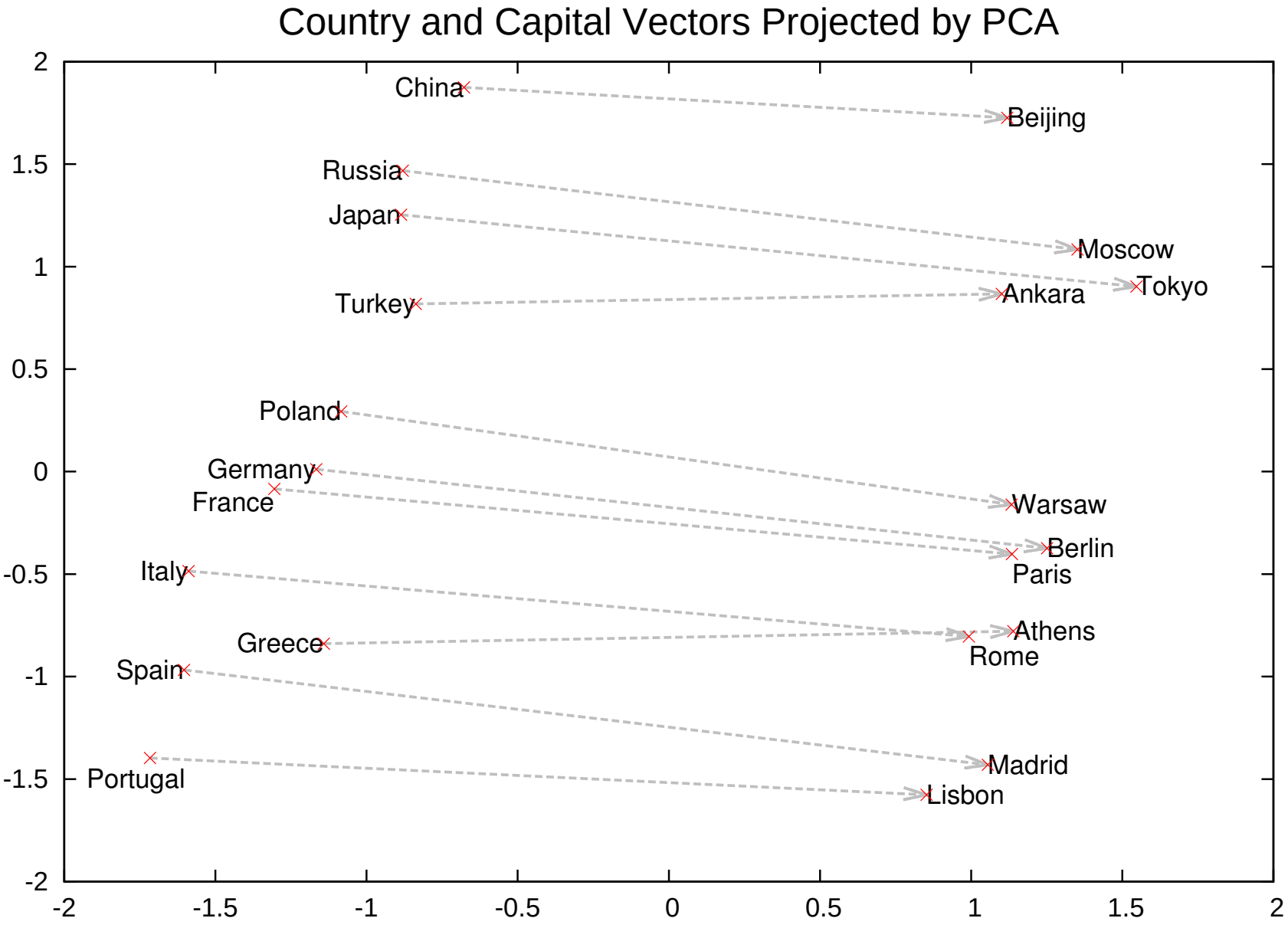
král - muž + žena = ?

král - muž + žena = královna

Tomáš Mikolov, 2012, word2vec

https://projector.tensorflow.org/

Word embeddings



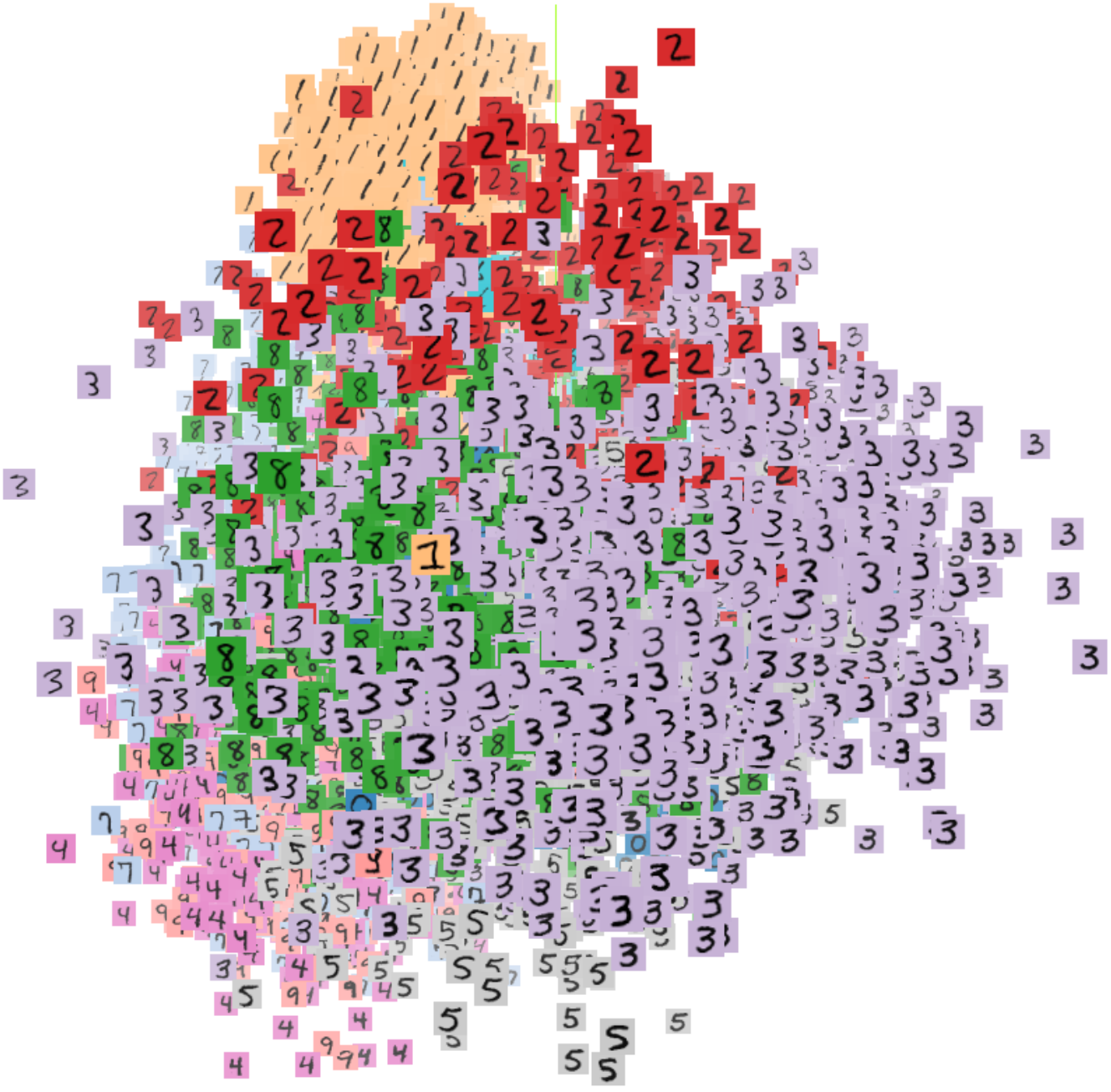
Word embeddings

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

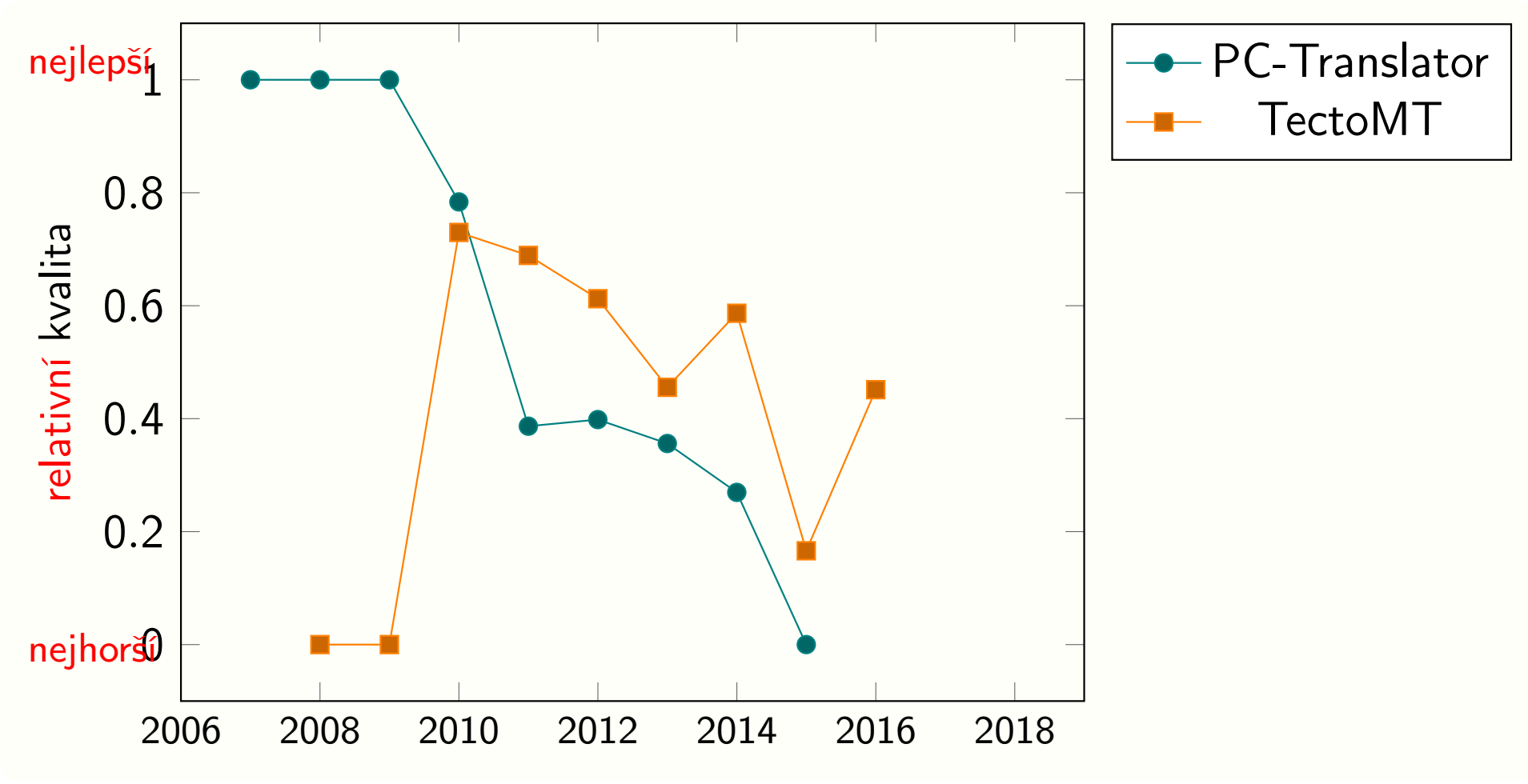
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

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

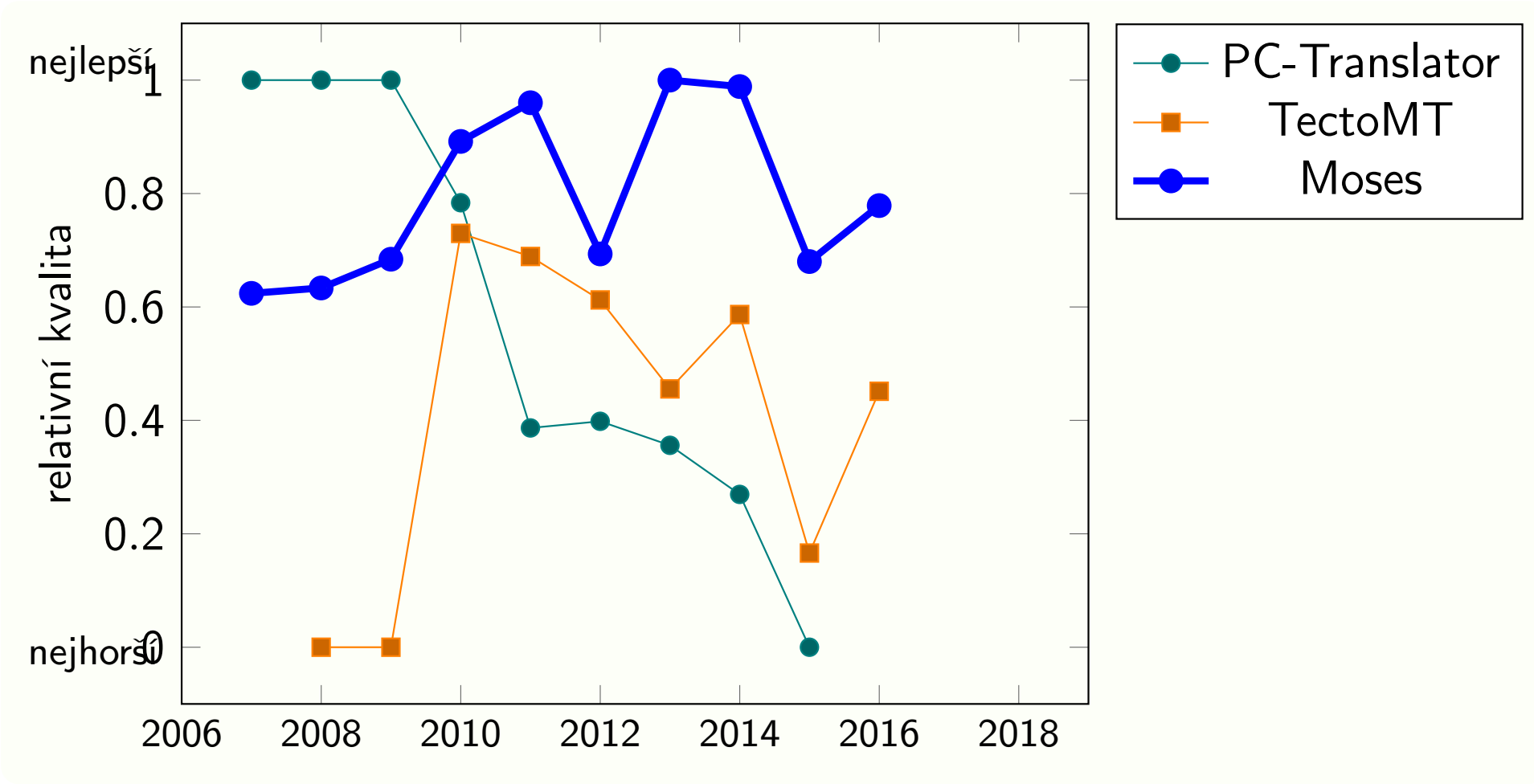
Word embeddings



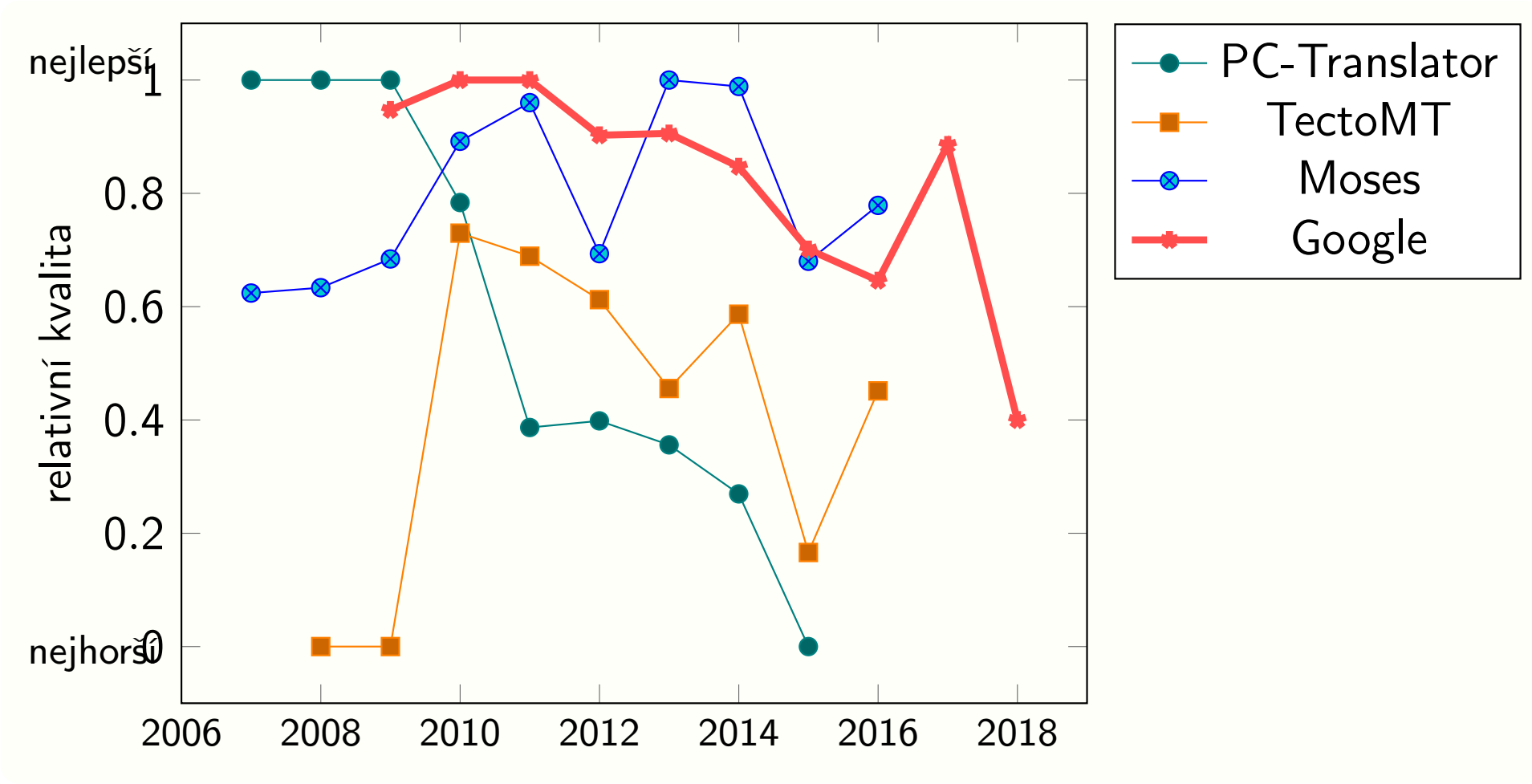
Strojový překlad angličtina→čeština 2007–2018



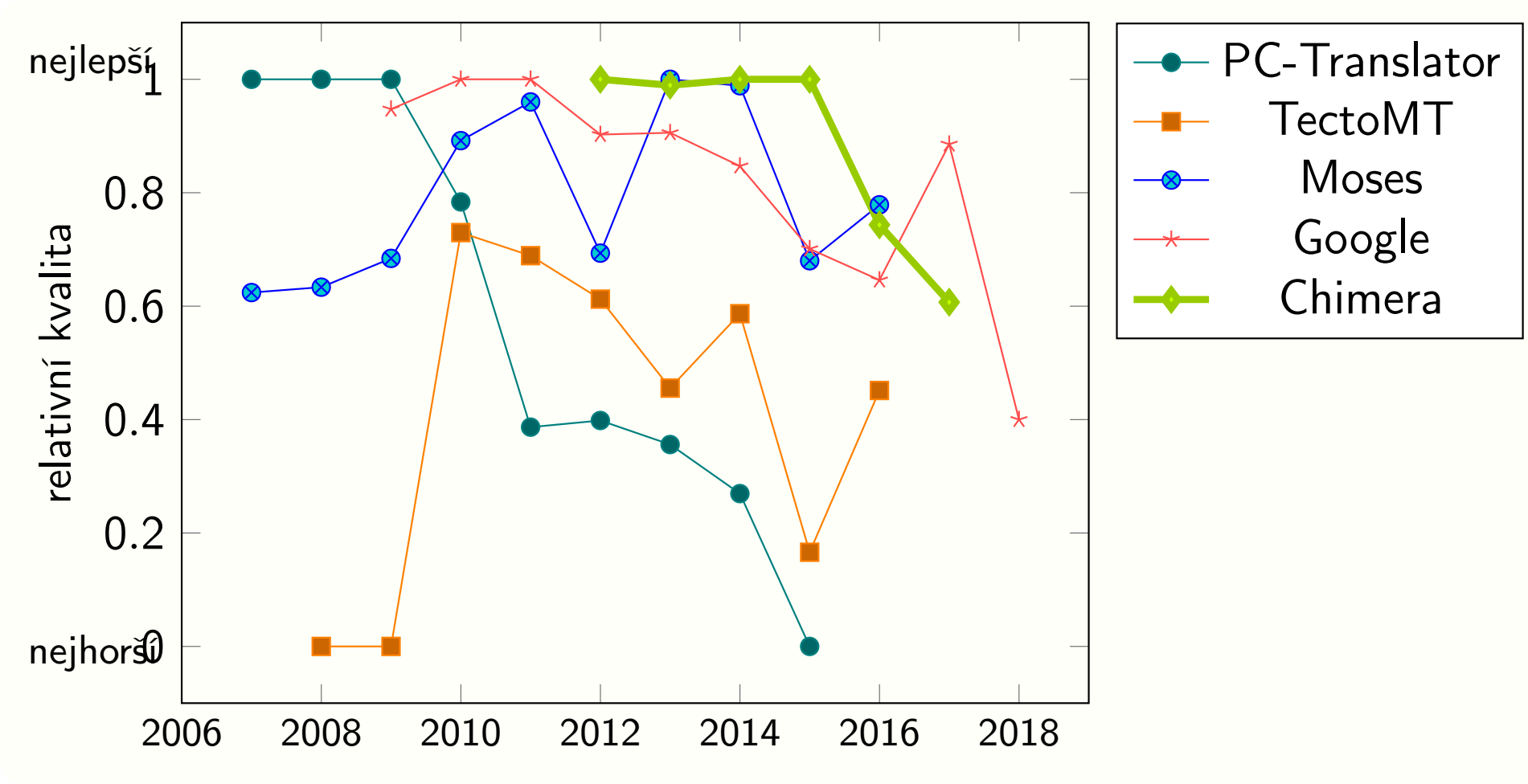
Strojový překlad angličtina→čeština 2007–2018



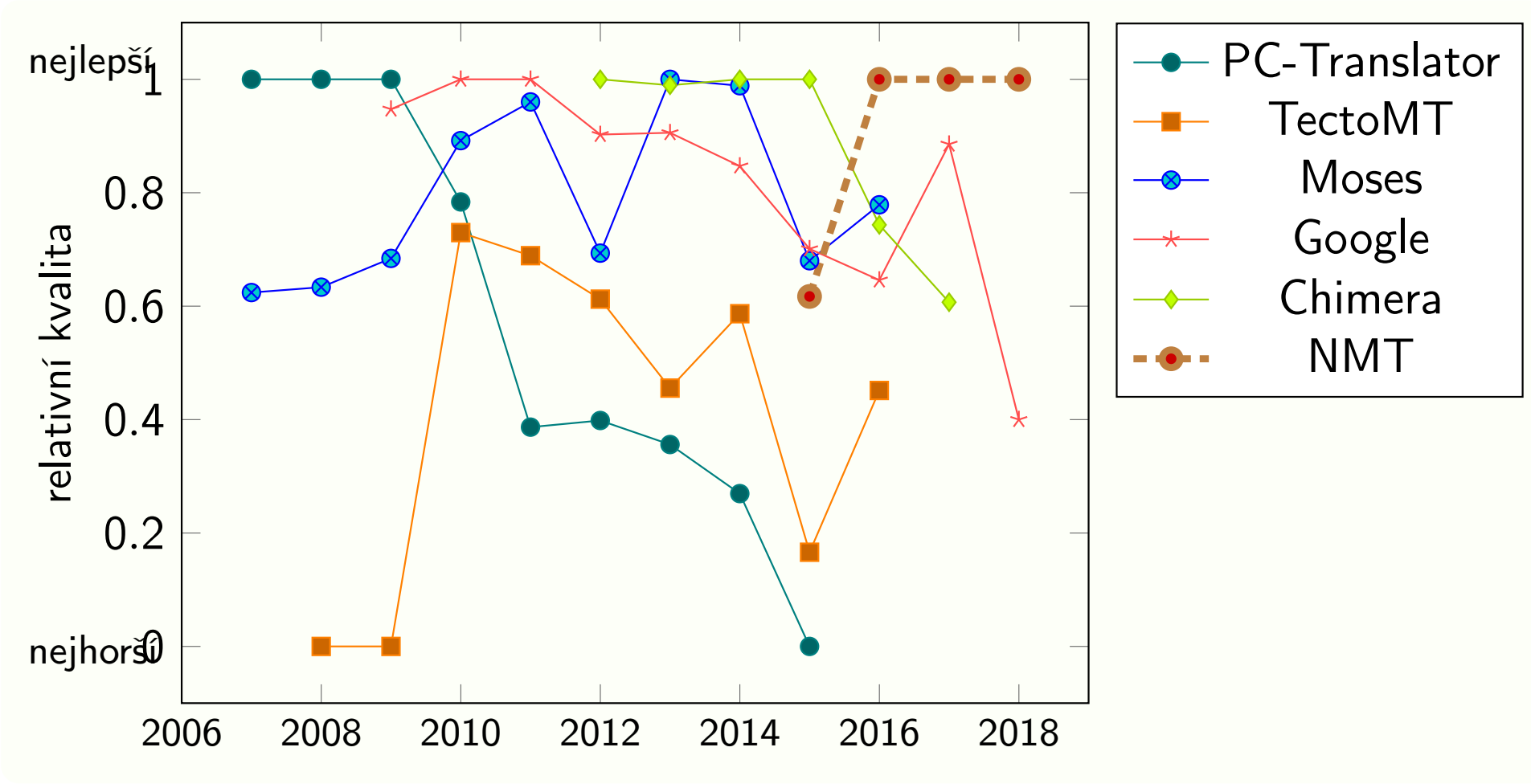
Strojový překlad angličtina→čeština 2007–2018



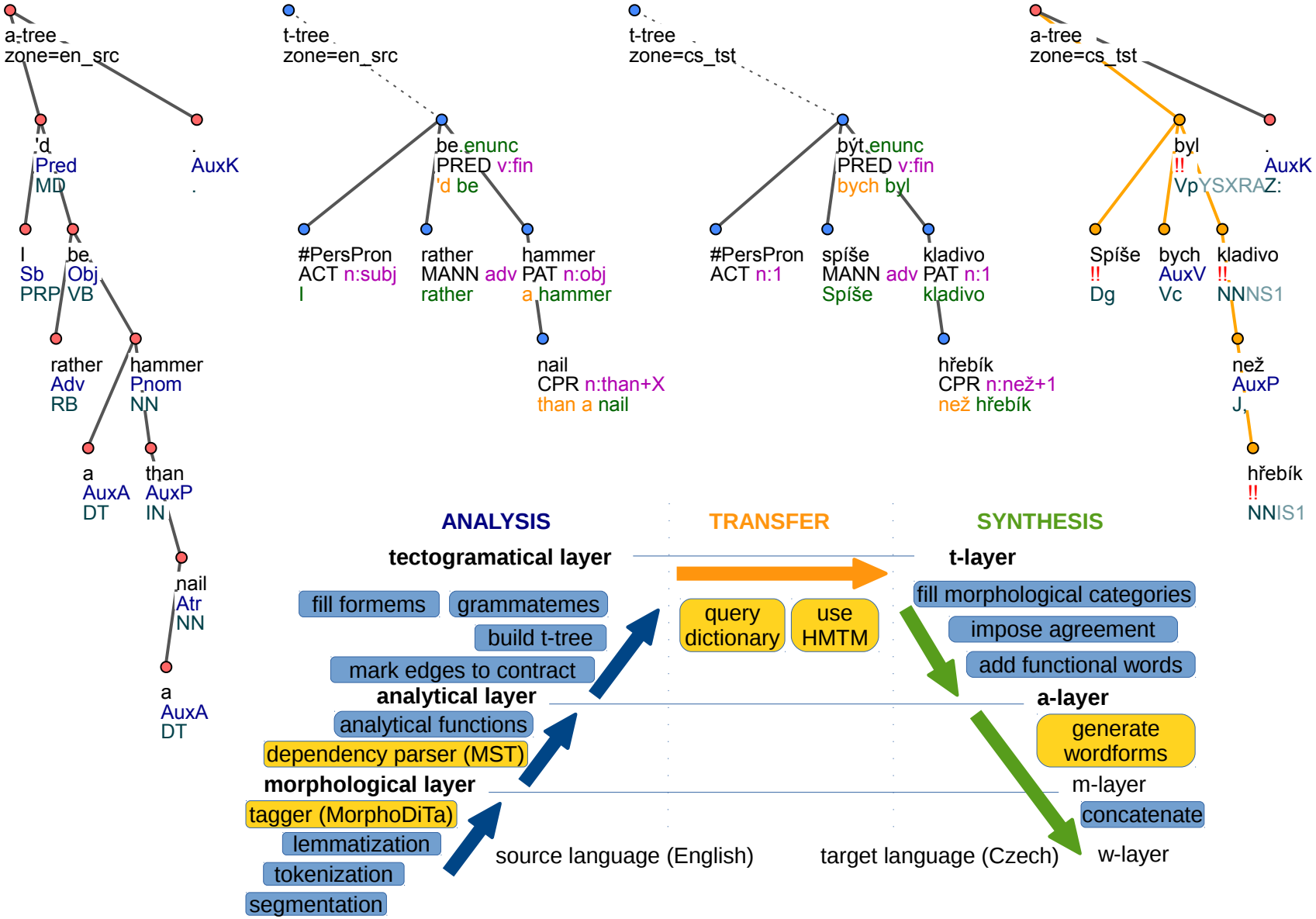
Strojový překlad angličtina→čeština 2007–2018



Strojový překlad angličtina→čeština 2007–2018



Hloubkově-syntaktický překladač TectoMT



I'd rather be a hammer than a nail.

Spíše bych byl kladivo než hřebík/nehet.

Strojové učení: rysy použité v překladovém modelu

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
...	

Strojové učení: rysy použité v překladovém modelu

output_label=hřebík#N

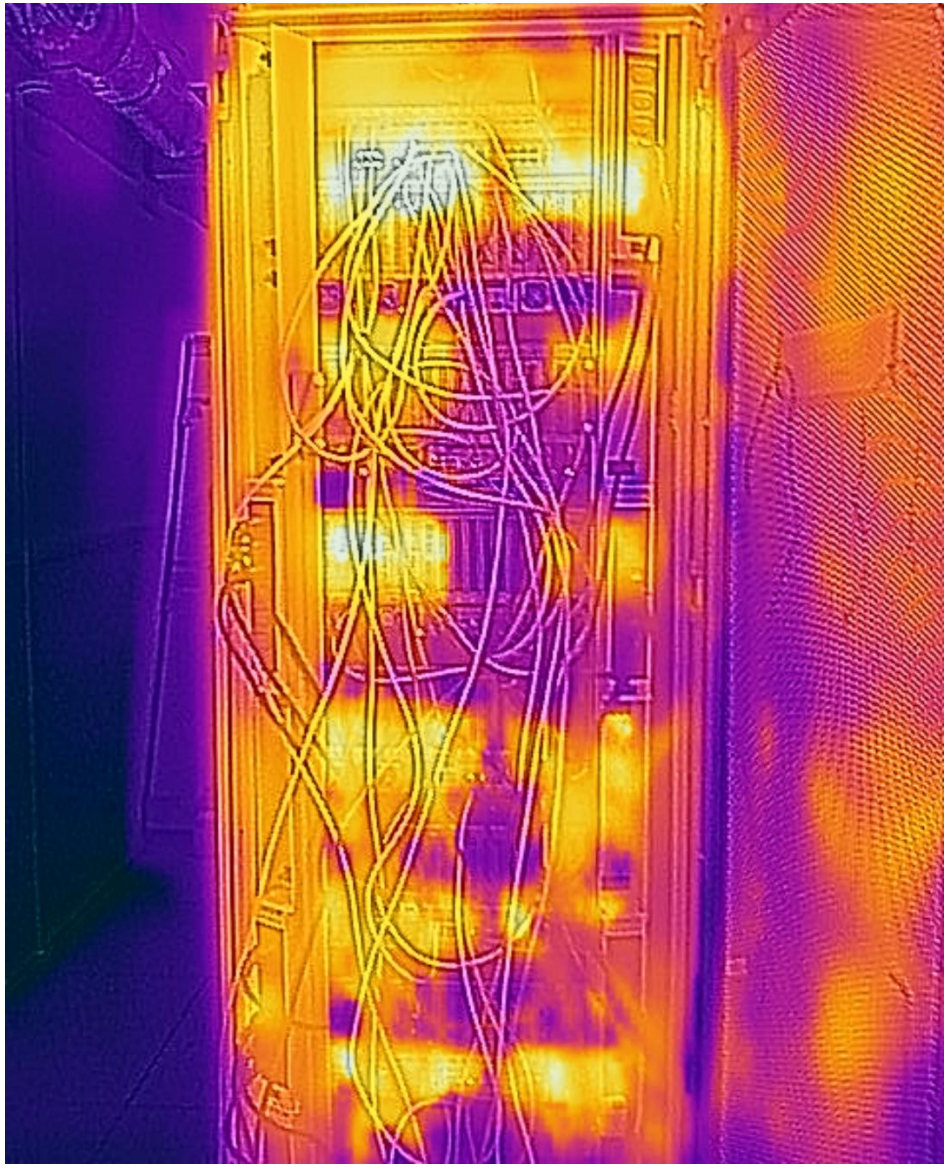
feature	λ
child_formeme_n:in+X=1	1.64
is_member_of_coord=1	1.30
child_formeme_v:fin=1	1.04
next_lemma=down	0.84
is_capitalized=1	0.79
+**precedes_parent=0**	**0.75**
tense_g=post	0.74
+**voice_g=active**	**0.66**
prev_lemma=drive	0.66
parent_capitalized=1	0.62
formeme=n:from+X	0.60
+**prev_lemma=hammer**	**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
...	

Co je to?

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

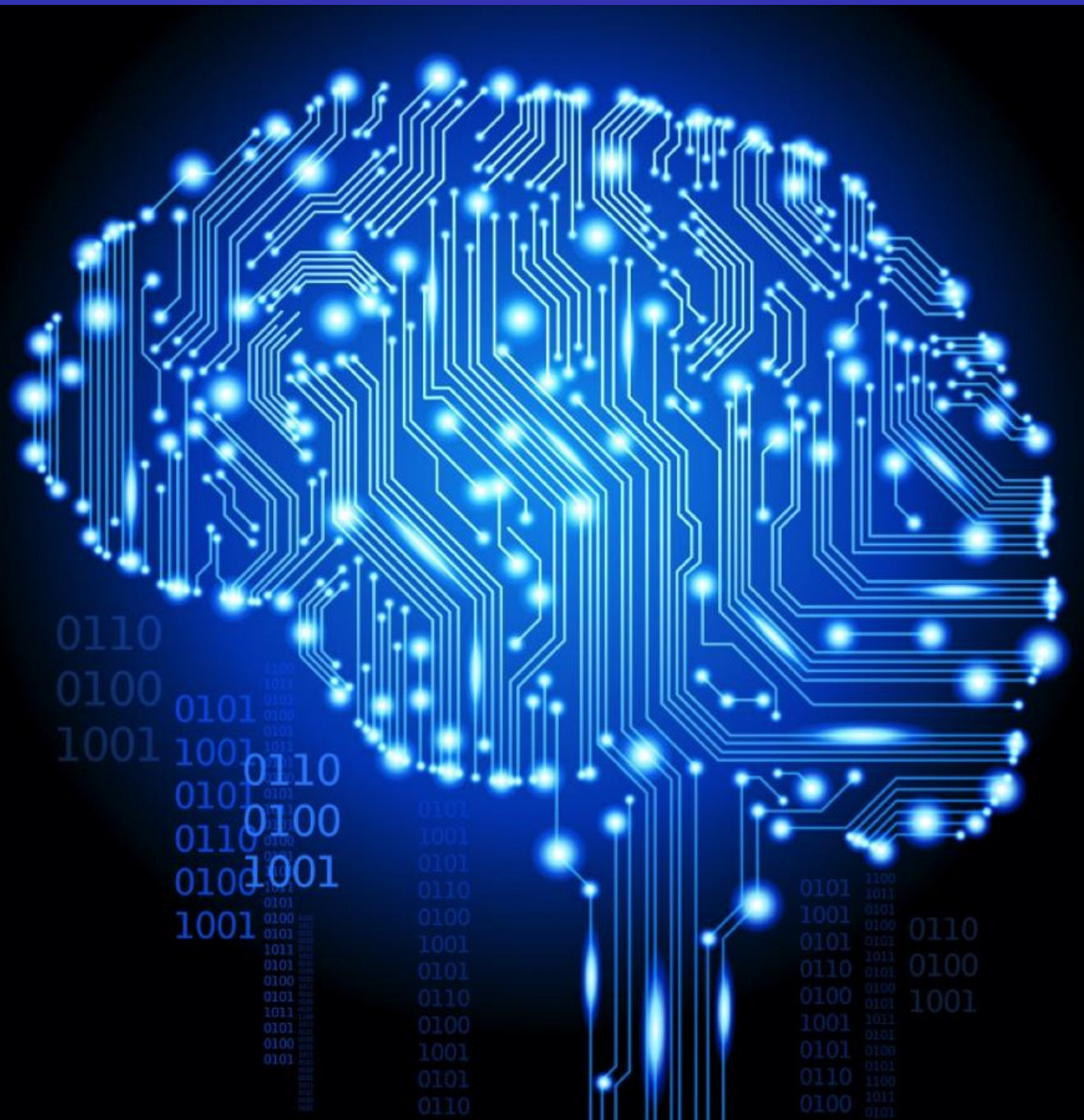


Umělá inteligence a její podobory/vývoj

umělá inteligence
∼1950

strojové učení
∼1980

hluboké učení
∼2010

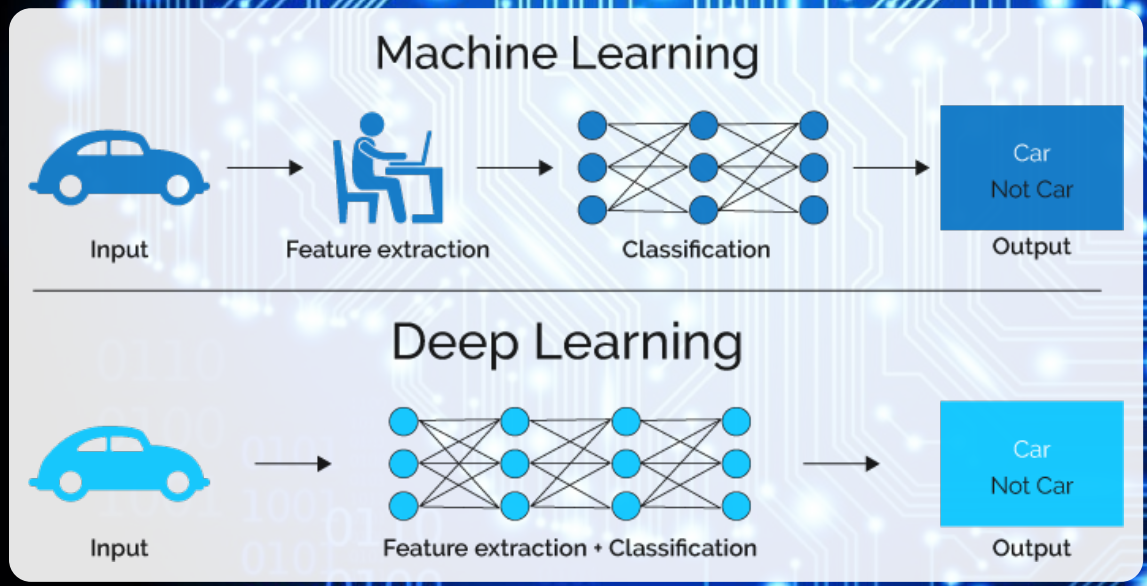


Umělá inteligence a její podobory/vývoj

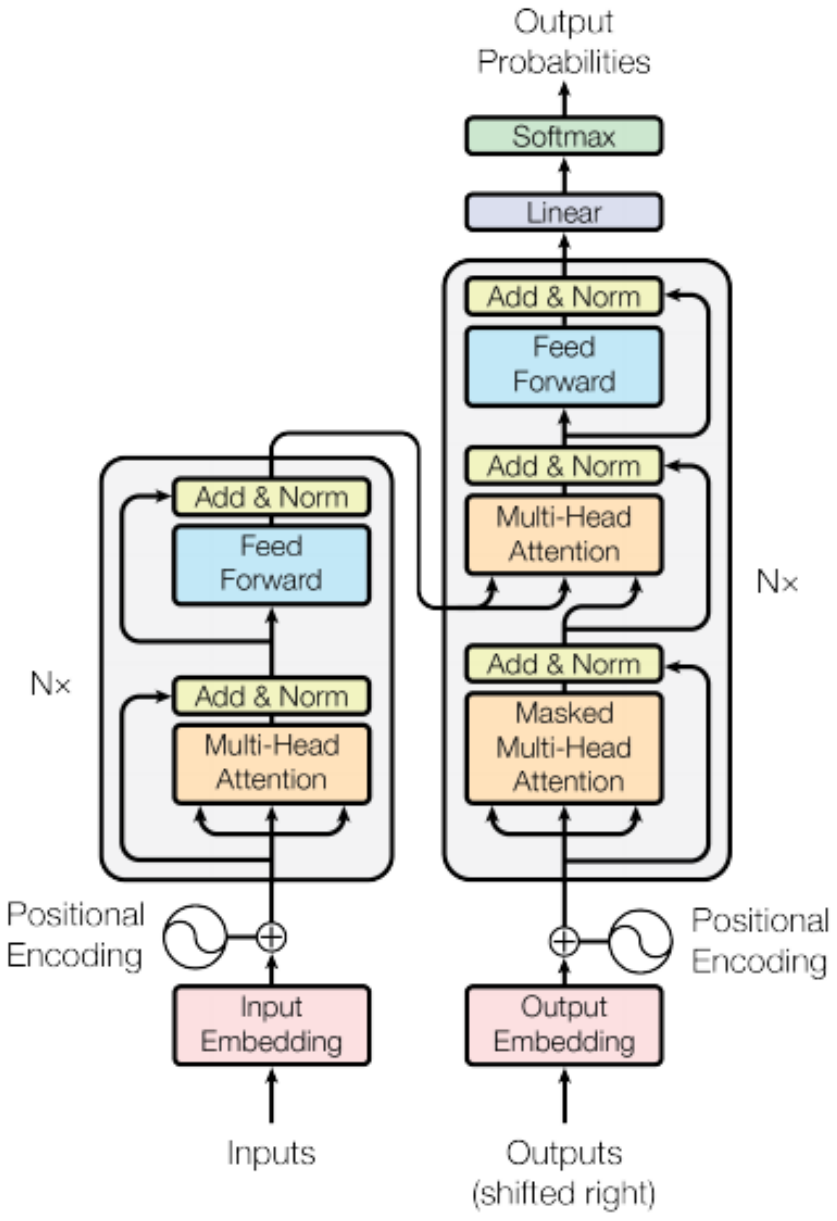
umělá inteligence
~1950

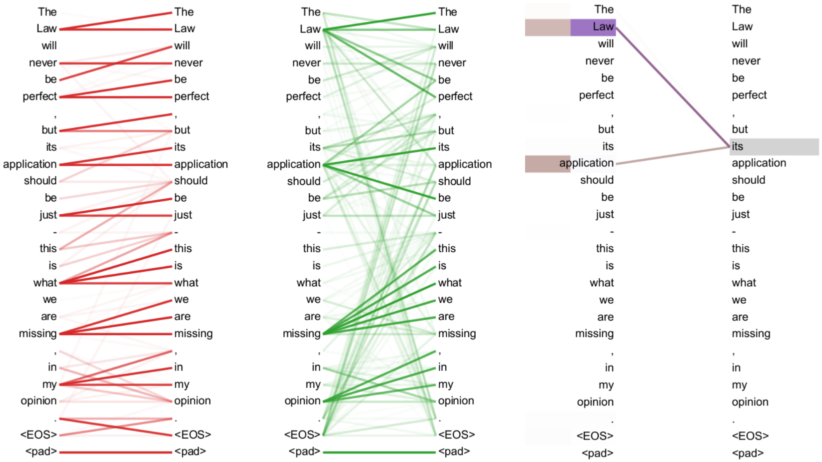
strojové učení
~1980

hluboké učení
~2010



Neuronový překladač Transformer




The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>
The Law will never be perfect , but its application should be just - this is what we are missing , in my opinion . <EOS> <pad>

Vyhodnocení anglicko-českého překladu (WMT 2018)

	Ave. %	System
1	**84.4**	Transformer
2	79.8	Edinburgh
	78.6	člověk
4	68.1	Google
5	59.4	Bing
6	54.1	Yandex

Rizika strojového překladu

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

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.
Transformer	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.
Transformer	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 Transformer sami:

http://lindat.cz/services/transformer/