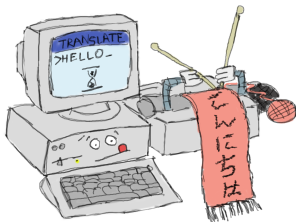


Strojový překlad a umělá inteligence

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 vysíláčky jsou malí činitelé.
Google	Velcí mluvčí jsou malí lidé.
TectoMT	Velcí řečníci jsou malí vrazi.
Transformer	Velcí mluvkové jsou malí dřiči.

Čím se zabývá počítačová lingvistika?



Chlapci šly.

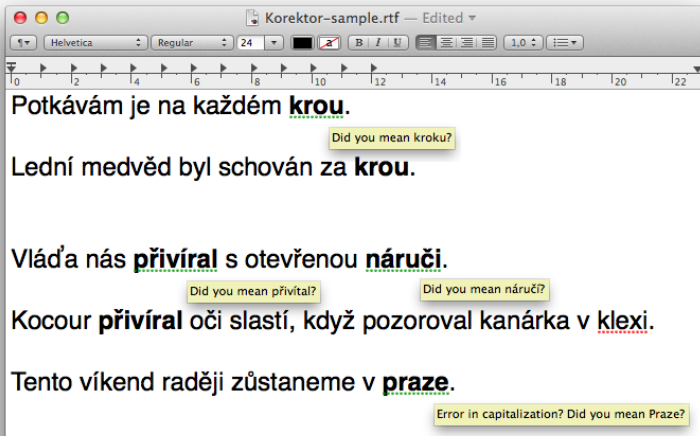
Chlapec šli do školy.

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

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

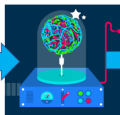


<http://ufal.cz/korektor>

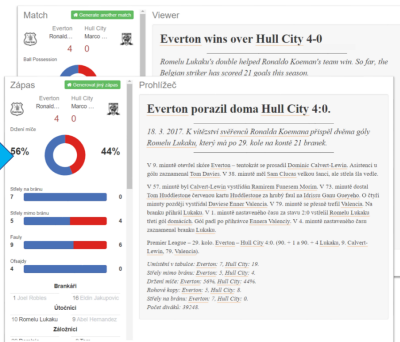


zpráv přibývá (FB, Twitter), novinářů ubývá, zisky klesají, čtenáři chtějí zprávy personalizované a hned

Date	Time	Match	Home	Away	Score	Goalkeeper
2017-08-11 17:00:00	17:00	Everton	Chelsea	0-0	Asmir Begović	
2017-08-11 19:00:00	19:00	Manchester City	Manchester United	1-0	David de Gea	
2017-08-11 20:00:00	20:00	Liverpool	Manchester City	0-0	Simon Mignolet	
2017-08-12 12:00:00	12:00	Manchester United	Manchester City	1-0	David de Gea	
2017-08-12 14:00:00	14:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-12 16:00:00	16:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-12 18:00:00	18:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-12 20:00:00	20:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-13 12:00:00	12:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-13 14:00:00	14:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-13 16:00:00	16:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-13 18:00:00	18:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-13 20:00:00	20:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-14 12:00:00	12:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-14 14:00:00	14:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-14 16:00:00	16:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-14 18:00:00	18:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-14 20:00:00	20:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-15 12:00:00	12:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-15 14:00:00	14:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-15 16:00:00	16:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-15 18:00:00	18:00	Manchester City	Manchester United	0-0	David de Gea	
2017-08-15 20:00:00	20:00	Manchester City	Manchester United	0-0	David de Gea	



Structured data
(real-time & historical)



Article / Report



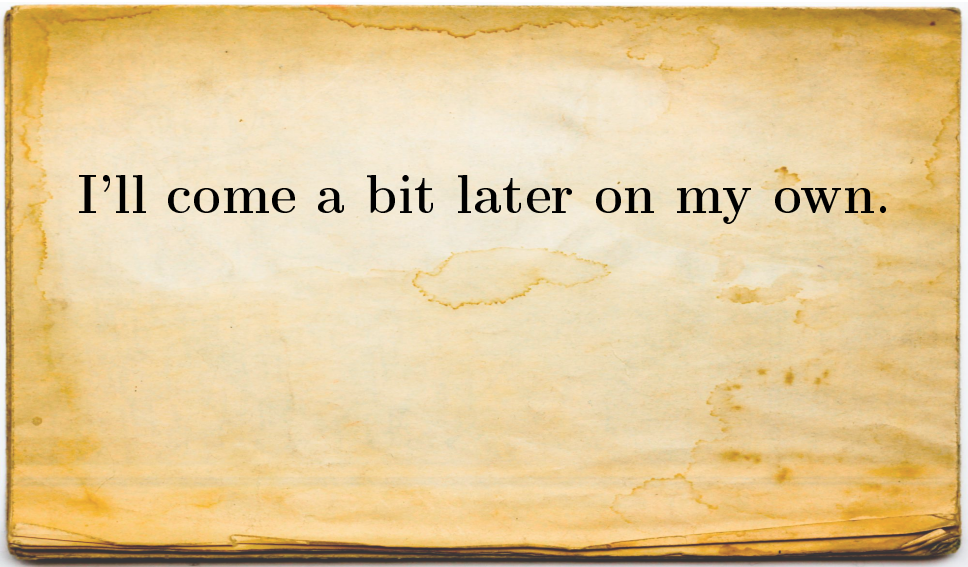
<http://quest.ms.mff.cuni.cz/verse/>

Byl by to rytíř, kde v pláně hřích vzlet,
Vědě jsem jse seheldo na přídoutně v světě si nezastává:
„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 vzítí,
na kónku je, milý svěžek.

A rectangular piece of aged, yellowish-tan paper with significant water damage and staining, particularly along the top and right edges. The paper has a mottled texture and some darker spots. Centered on the paper is a single line of text in a black, serif font.

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 „kapřící“

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

KATEŘINA KOLÁŘOVÁ

V e druhém patře moderní pražské kancelářské budovy Nile House usedá k jednacím stolu pětice 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“ dodavatelům služeb, a za úplatek mu umožnili vyhrát lukrativní zakázku. K vyšetřování používají specializované techniky 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 Kračinová, 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 prokávají izolovaně u prošetřované společnosti klienta. Proto často, paralelně s naším šetřením, probíhá i policejní vyšetřování,“ říká Kračinová.

Rychlý zášah



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 Kračinová (vpravo).

FOTO MAFRA – DAN MATERNA

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

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

Rozpoznání kódované řeči

„Na základě blízkosti slovní záso-

Důkazem manipulace s výsledky tendru je typicky vznik

mošla existovat, protože o vítězi tendru ještě nebylo nákupním oddělením oficiálně rozhodnuto,“ popisuje Kračinová. Pak už analytici hledají v počítačích konkrétní textový soubor.

Kateřina Veselovská

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

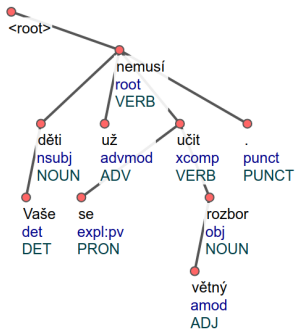
Jaroslava Kračinová

■ 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 právníků osob, corporate governance a ochranu osobních údajů.

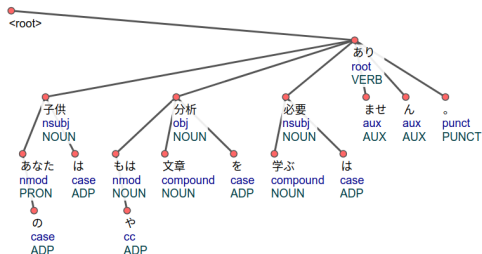
rozmolněné hranice toho, co je v této oblasti už za hranou a co je ještě v pořádku. V některých zemích je to ale něco naprosto nepřehledného,“ vzpomíná Kračinová. Byl tím specialistů prokázal, že se

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

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。



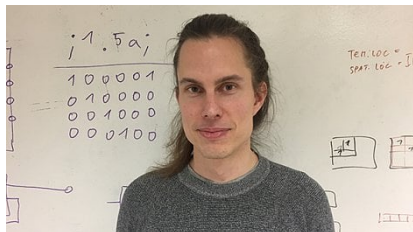
<http://lindat.cz/services/udpipes/>

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

král - muž + žena = ?

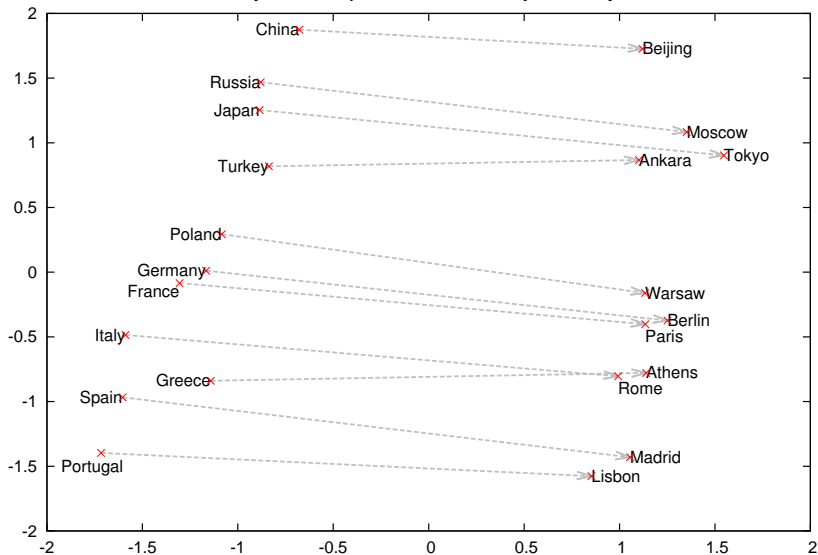
král - muž + žena = **královna**

Tomáš Mikolov, 2012, word2vec



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

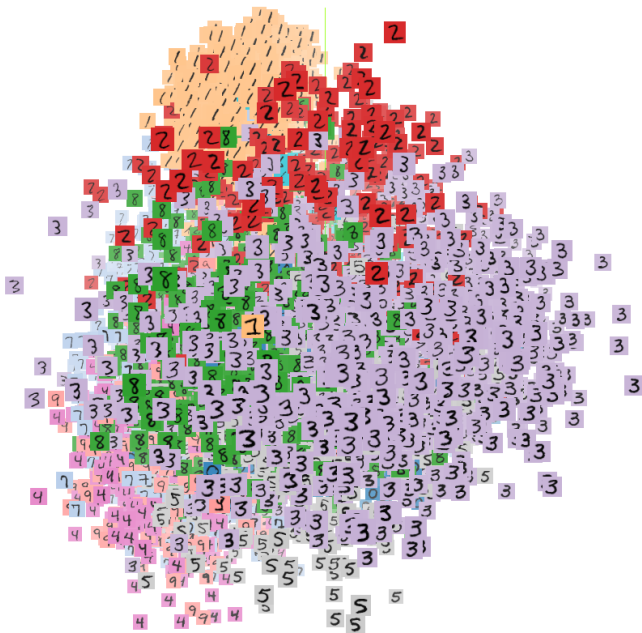
Country and Capital Vectors Projected by PCA

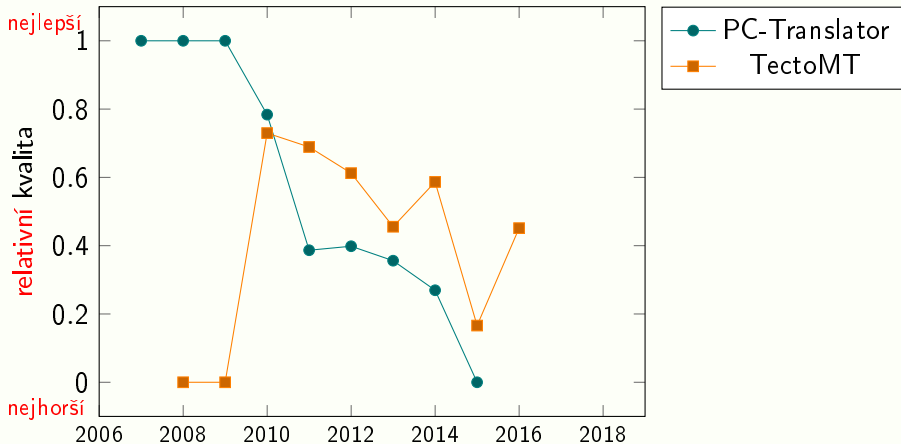


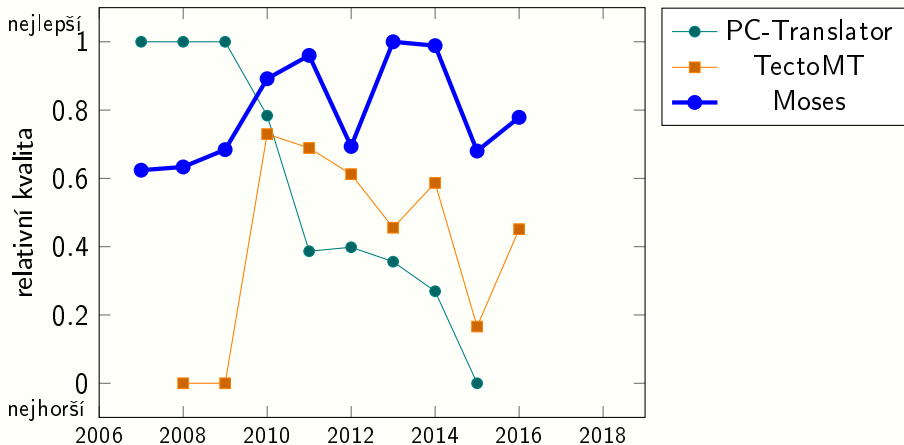
Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna Check crown Polish zolty CTK	Hanoi Ho Chi Minh City Viet Nam Vietnamese	airline Lufthansa carrier Lufthansa flag carrier Lufthansa Lufthansa	Moscow Volga River upriver Russia	Juliette Binoche Vanessa Paradis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

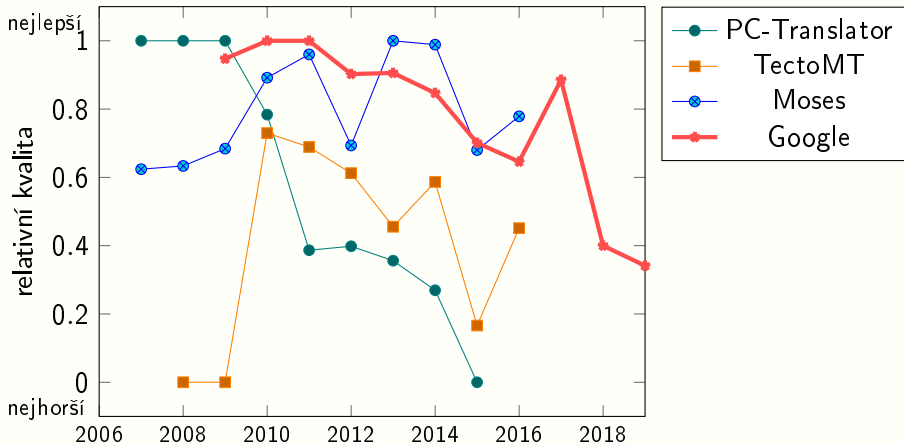
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

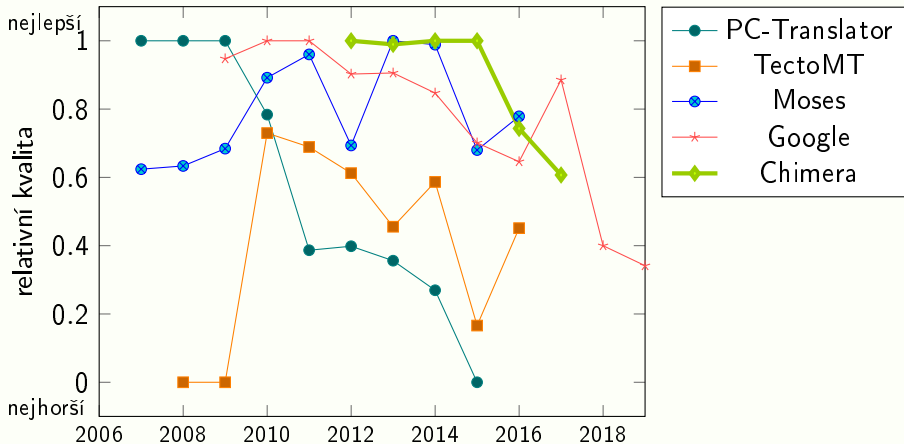
Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - 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

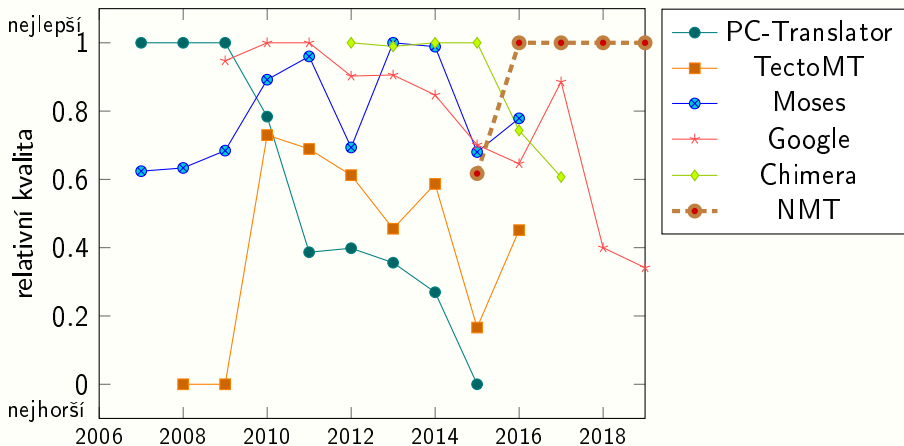


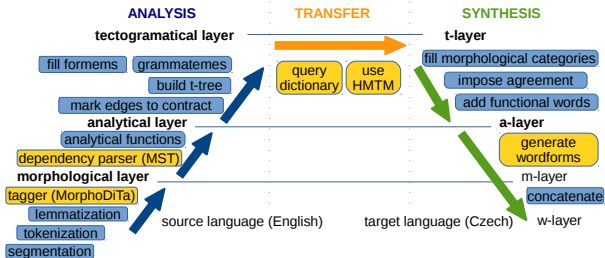
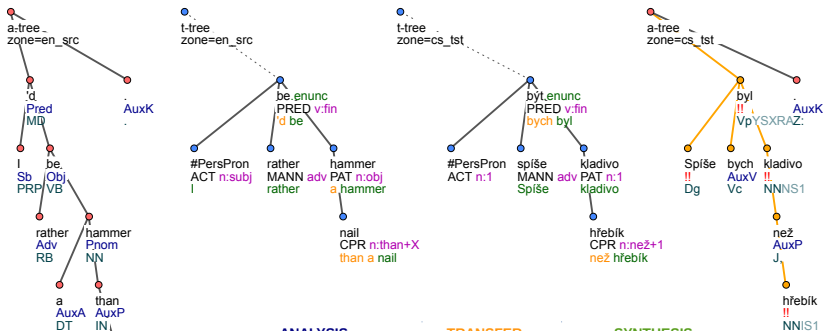












I'd rather be a hammer than a nail.

Spíše bych byl kladivo než hřebík/nehet.

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=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
...	

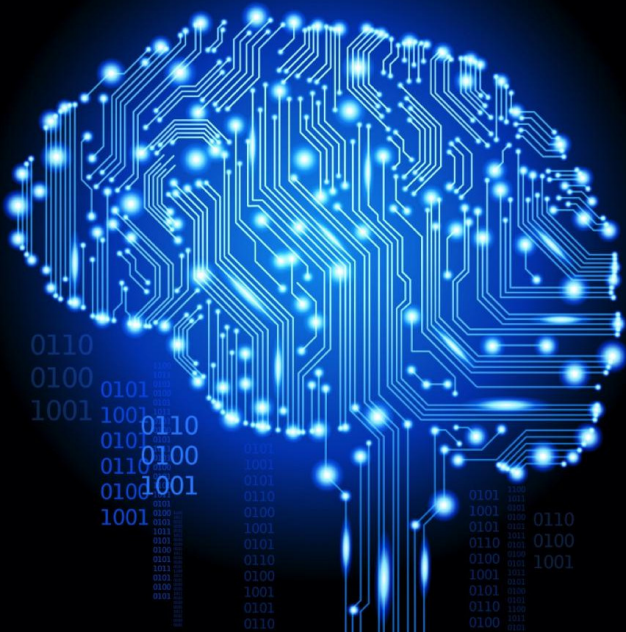




umělá inteligence
~1950

strojové učení
~1980

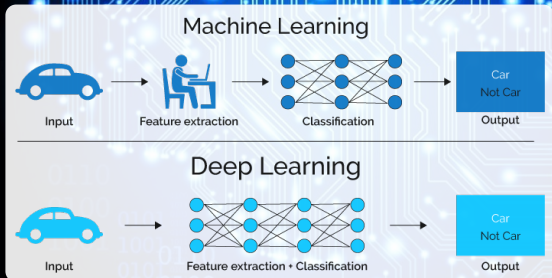
hluboké učení
~2010

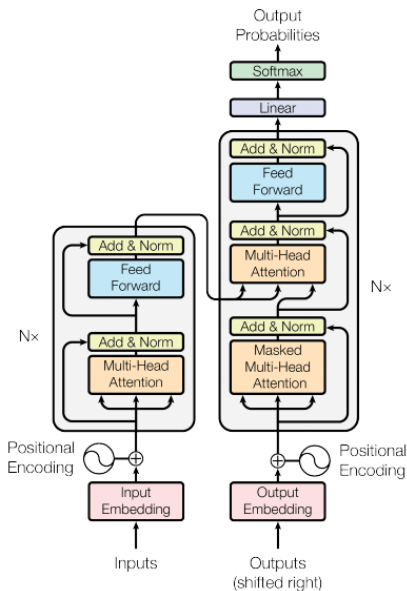


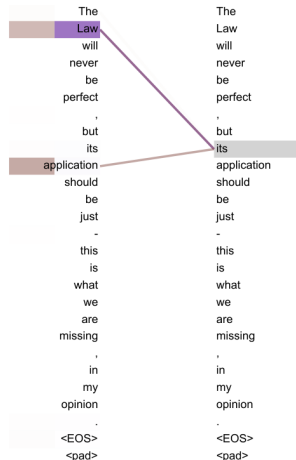
umělá inteligence
~1950

strojové učení
~1980

hluboké učení
~2010







	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

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



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 pěřového hejna spolu.
Vrána k vráně sedá.

Zkuste si Transformer sami:

<http://lindat.cz/services/translation/>