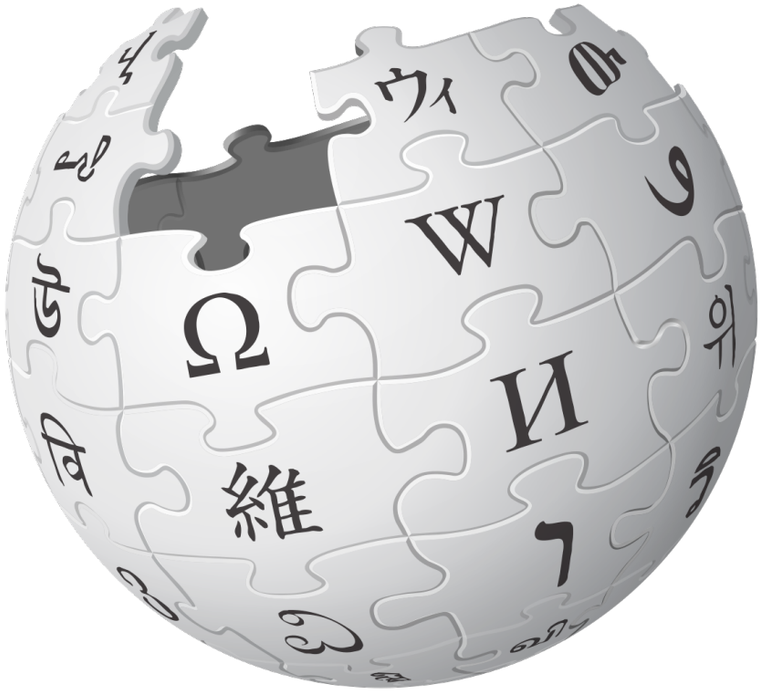
Wikipedie



z pohledu

korpusové lingvistiky

Marek Blahuš,
XI. Wikikonference
Pardubice, 23.11.2019

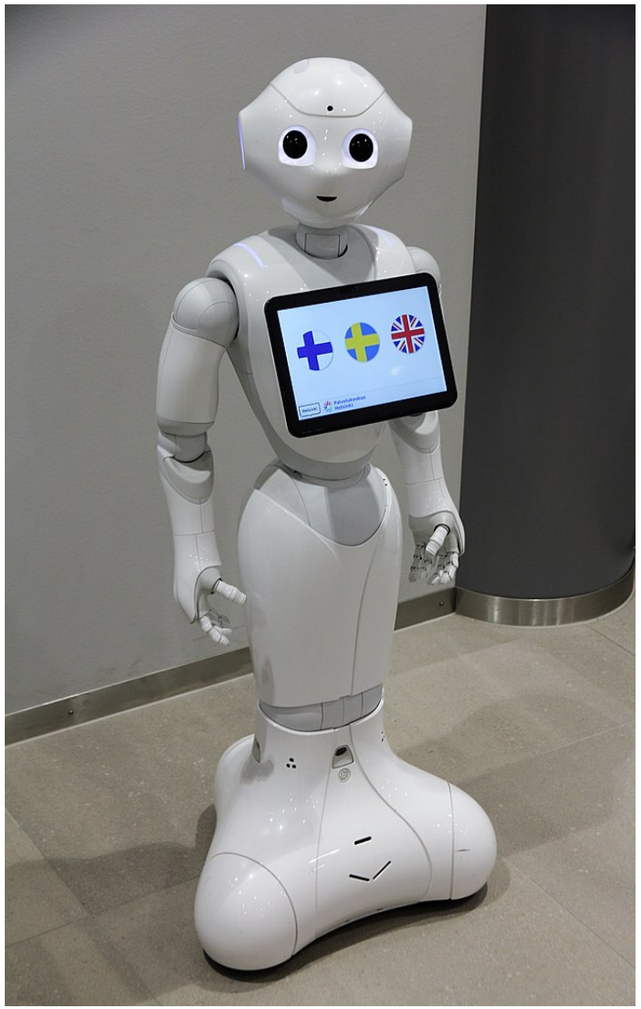
Marek Blahuš

- moderátor Wikikonferencí 2017–19
- wikipedista od roku 2003 (Blahma)
- člen spolku Wikimedia ČR
- absolvent FI MU (počítačová lingvistika)
- informatik a polyglot
- programování a zpracování dat
- od roku 2017 Lexical Computing CZ, Brno

Počítačová lingvistika

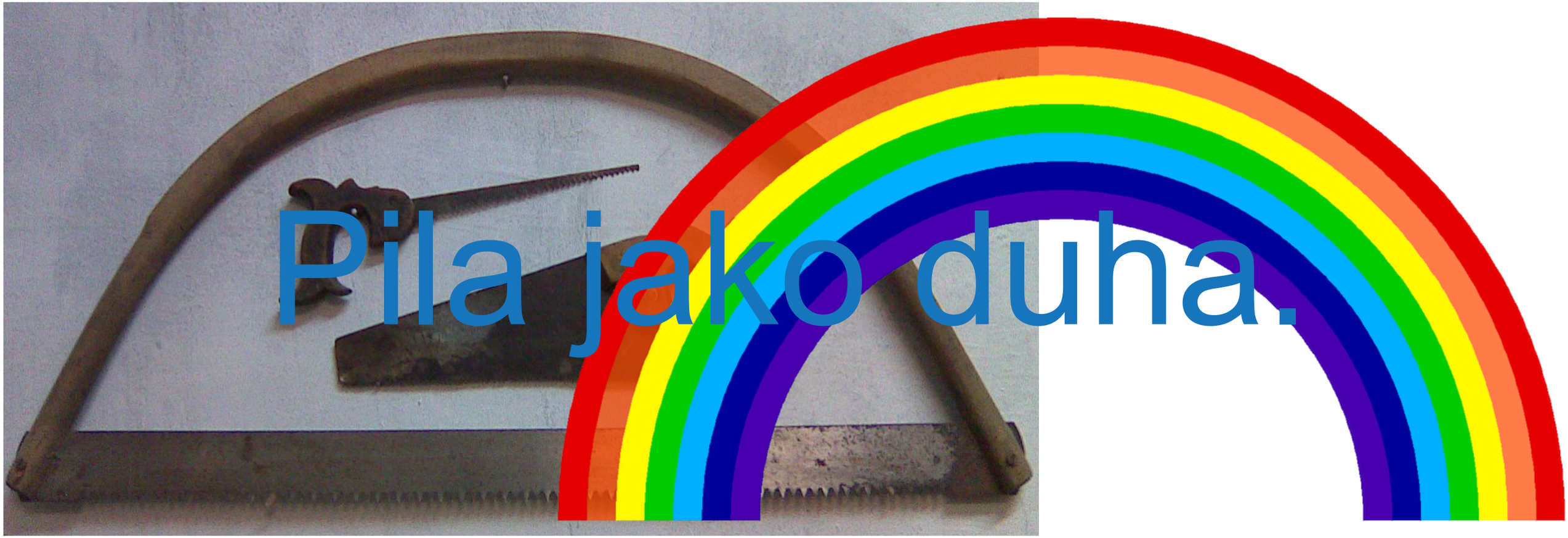
(Zpracování přirozeného jazyka, NLP)

- Informatika + Lingvistika
- Porozumění lidskému jazyku strojem
- Aplikace:
  - dialogové systémy
  - strojový překlad
  - extrakce informací
  - korektura textu
  - prediktivní psaní
  - určení autorství



Kolik významů mohou mít tyto věty?

Pila jako duha.



https://commons.wikimedia.org/wiki/File:Stare_pily.jpg – Mnemosine, PD
https://commons.wikimedia.org/wiki/File:Rainbow-diagram-ROYGBIV.PNG – ProfDEH, PD

Splácej to,
jak umíš.

Praštil se sluchátkem.

Praštil se sluchátkem.

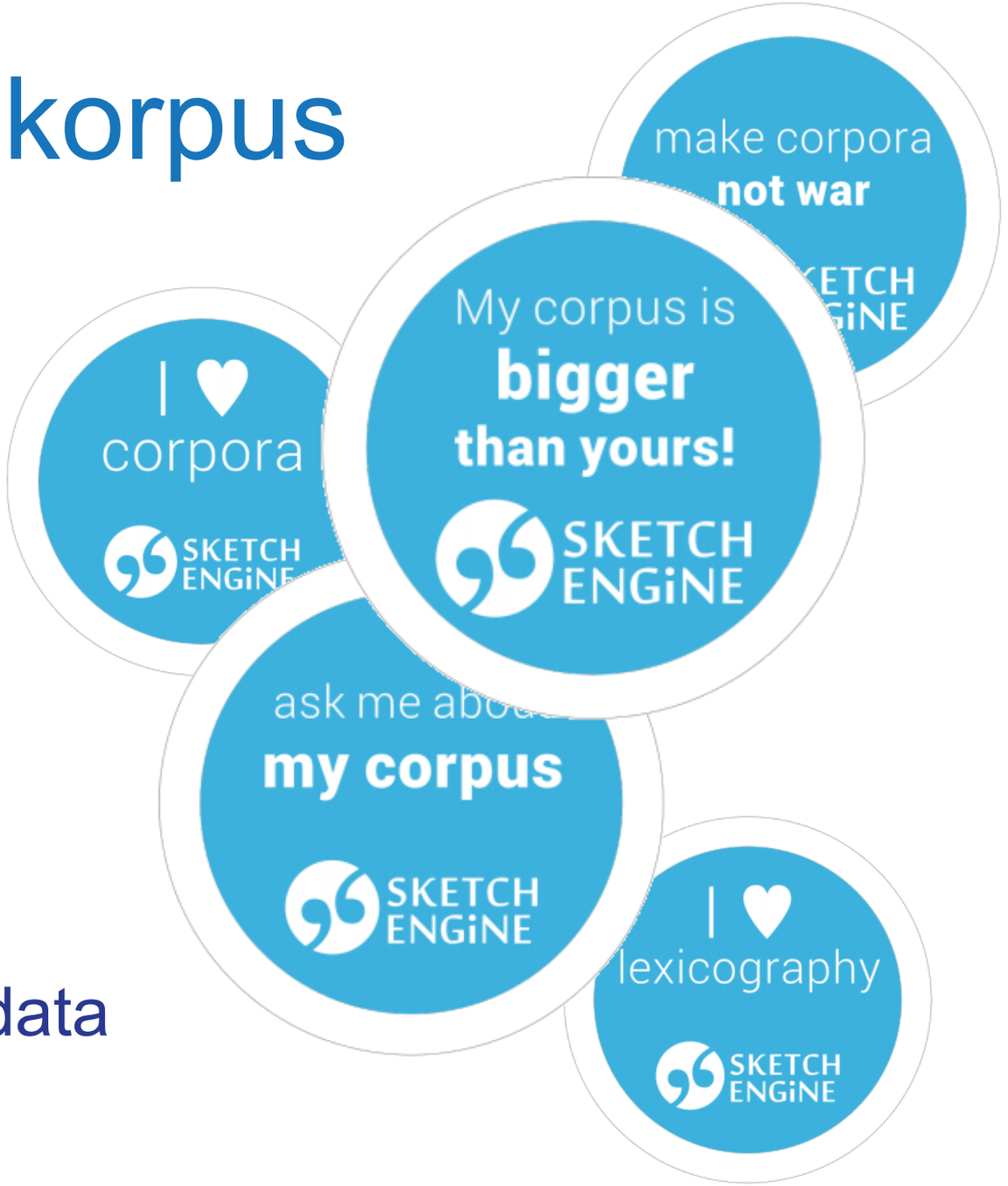
https://commons.wikimedia.org/wiki/File:De_broertjes_Thomas_en_Holger,_Bestanddeelnr_254-3608.jpg – Willem van de Poll, CC0

Bál se hrabat
v starých hadrech.

Bál se hrabat
v starých hadrech.

Jazykový korpus

- Rozsáhlá databáze textů
- Vzorek jazyka v akci
- Korpusový manažer (= program)
  - Sketch Engine (od roku 2003)
  - www.sketchengine.eu
- Jazyková preskripce vs. deskripce
- Počítačová lexikografie = slovníky
- Velikost, vyváženost, původ, metadata



CONCORDANCE

Czech Web 2017 (csTenTen17)

simple jablko **221,573** > sample 200 **200** (0.02 per million)

KWIC

	Details	Left context	KWIC	Right context
1	simplysay.cz	laterna - Granátové :erna/k1gFnSc1 -/kIx~ granátový/k2eAgNnSc1d1	**jablko** jablko/k1gNnSc1	a vlašské ořechy </ a/k8xC vlašský/k2eAgInPc1d1 ořech/k1gInPc1
2	recepty.cz	ořechy a štávu z h/k1gInPc1 a/k8xC štávat/k5eAaPmIp1nS z/k7c2	**jablek** jablko/k1gNnPc2	jsem moc nemačkala být/k5eAaImIp1nS moc/k6eAd1 mačkat/k5eNaImA
3	zlate-mince.cz	udolfa II . - žezlo , gMnSc2 II/kA ./kIx. -/kIx~ žezlo/k1gNnSc4 ,/kIx,	**jablko** jablko/k1gNnSc4	a korunu . </s><s> Zpra a/k8xC koruna/k1gFnSc4 ./kIx. zprav
4	energiezivota.com	chrup vtisknutý do :hrup/k1gInSc4 vtisknutý/k2eAgInSc4d1 do/k7c2	**jablka** jablko/k1gNnPc1	Krásná jak stromy v krásný/k2eAgNnPc1d1 jak/k8xC strom/k1gInPc1 v
5	byznys.lidovky.cz	masivní dovozy italských gInPc1d1 dovoz/k1gInPc1 italský/k2eAgNnPc2d1	**jablek** jablko/k1gNnPc2	. </s><s> Z Francie se ./kIx. z/k7c2 Francie/k1gFnSc2 se/k3
6	jaktak.cz	) . </s><s> Nastrouhaná) ./kIx. nastrouhaný/k2eAgNnPc1d1	**jablka** jablko/k1gNnPc1	si rozdělte na 3 se/k3xPyFc3 rozdělit/k5eAaPmRp2nP na/k7c4 [nu
7	iforum.cuni.cz	<s> A pořekadlo , že a/k8xC pořekadlo/k1gNnSc4 ,/kIx, že/k8xS	**jablko** jablko/k1gNnSc1	nepadá daleko od stro padat/k5eNaImIp3nS daleko/k6eAd1 od/k7c2 stro
8	mojezdravi.cz	, brambory , grepy , lx, brambor/k1gInPc1 ,/kIx, grep/k1gInPc1 ,/kIx,	**jablka** jablko/k1gNnPc1	, libové maso , ,/kIx, libový/k2eAgNnSc4d1 maso/k1gNnSc4 ,/kIx
9	mojebetynka.ma...	;><s> Mezitím oloupeme mezitím/k6eAd1 oloupat/k5eAaPmIp1nP	**jablko** jablko/k1gNnSc4	, vykrojíme jádřinec a ,/kIx, vykrojit/k5eAaPmIp1nP jádřinec/k1gInSc4 a/
10	recenze-hotelu.i...	tam byly hrušky ,	**jablka** 	, meruňky a málo

WORD SKETCH

Czech Web 2017 (csTenTen17)

jablko as noun 221,573×

prepositional phrases | verbs with "jablko" as locale object | "jablko" is ...

modifiers of "jablko"

- **granátový** — granátového jablka
- **nastrouhaný** — nastrouhaná jablka
- **nakousnutý** — s nakousnutým jablkem
- **oloupaný** — oloupaná jablka
- **strouhaný** — strouhaná jablka
- **nakrájený**

... of "jablko"

- **odrůda** — odrůd jablek
- **úroda** — úrodu jablek
- **plátek** — plátky jablek
- **sklizeň** — sklizeň jablek
- **kilo** — kilo jablek
- **vůně**

verbs with "jablko" as subject

- **padat** — že jablko nepadá daleko od stromu
- **změknout** — dokud jablka nezměknou
- **rozvařit** — se jablka rozvaří
- **padnout** — jablko nepadne daleko od stromu
- **dozrávat** — jablka dozrávají

verbs with "jablko" as accusative object

- **nastrouhat** — nastrouháme jablka
- **oloupat** — oloupeme jablka
- **míchat** — míchá jablka
- **sníst** — sníst jablko
- **utrhnout** — utrhla jablko
- **rozkrojit**

"jablko" and/or ...

- **hruška** — jablka a hrušky
- **mrkev** — jablka a mrkev
- **banán** — jablka a banány
- **pomeranč** — jablka a pomeranče
- **skořice** — jablky a skořicí
- **žezlo**

English • German • Russian • Czech • French • Spanish • Japanese • Polish • Arabic • Italian • Catalan • Portuguese • Turkish • Swedish • Hungarian • Romanian • Dutch • Ukrainian • Danish • Chinese Simplified • Chinese Traditional • Greek • Norwegian (Mixed) • Finnish • Croatian • Norwegian Bokmål • Estonian • Slovak • Hebrew • Afrikaans • Albanian • Amharic • Azerbaijani • Basque • Belarusian • Bengali • Bosnian • Bulgarian • Cantonese • Cuneeee Wangka • Filipino • Frisian • Georgian • Gujarati • Hausa (Boko) • Hindi • Icelandic • Igbo • Indonesian • Irish • Kannada • Kazakh • Korean • Kyrgyz • Latin • Latvian • Lithuanian • Macedonian • Malay • Malayalam • Maltese • Maori • Mongolian • Montenegrin • Nepali • NilKo • Norwegian Nynorsk • Oromo • Persian • Punjabi (Shahmukhi) • Samoan • Scottish Gaelic • Serbian • Serbian (Latin) • Setswana • Slovenian • Somali • Swahili • Tajik • Tamil • Tatar • Telugu • Thai • Tibetan • Tigrinya • Turkmen • Urdu • Uzbek • Vietnamese • Welsh • Yoruba • Ancient Greek • Armenian • Breton • Burmese • Esperanto • Galician • Kalaallisut • Kurdish (Sorani) • Kuwarra • Limburgish • Maduwongga • Maldivian • Mankulatjara • Mayjiljar • Marathi • Marlpa • Mirning • Ndebele • Newspeak • Ngaanyatjarra • Ngaju • Ngalia • Nganta • Northern Sotho • Nyakinyaki • Pashto • Pintupi • Pitjantjatjara • Quechua • Sanskrit (romanised) • Sesotho • Sinhalese • Swazi • Syriac • Tagalog • Talysh • Tjalkatjarra • Tjupan • Tsonga • Venda • Wangkatja • Warlpiri • Warlpiri • Wudjaarri • Xhosa • Yankunytjatjara • Yiddish • Zulu

>400 korpusů

>90 jazyků

$>10^{10}$ slov

Wikipedie pomáhá lingvistům.

Zná tisíce jazyků…

Writing system [edit | edit source]

Main article: Lao script

The Lao script, derived from the Khmer alphabet of the Khmer Empire in the 14th century,[13] is ultimately rooted in the Pallava script of South India, one of the Brahmi scripts.[14] Although the Lao script bears resemblance to Thai, the former contains fewer letters than Thai because by 1960 it was simplified to be fairly *phonemic*, whereas Thai maintains many etymological spellings that are pronounced the same.[15]

The script is traditionally classified as an abugida, but Lao consonant letters are conceived of as simply representing the consonant sound, rather than a syllable with an inherent vowel.[15] Vowels are written as diacritic marks and can be placed above, below, in front of, or behind consonants. The script also contains distinct symbols for numerals, although Arabic numerals are more commonly used.

The Lao liturgical script is written in the Tai Tham script[14] and is still used in temples in Laos and Isan.

Punctuation [edit | edit source]

Lao is traditionally not written with spaces between words, although signs of change are multiplying. Spaces are reserved for ends of clauses or sentences. Periods are not used, and questions can be determined by question words in a sentence. Traditional punctuation marks include ໌, an obsolete mark indicating silenced consonants; ๆ, used to indicate repetition of the preceding word; ๏, the Lao ellipsis that is also used to indicate omission of words; ฯ, a more or less obsolete symbol indicating shortened form of a phrase (such as royal names); and ฯลฯ, used to indicate *et cetera*.

In more contemporary writing, punctuation marks are borrowed from French, such as exclamation point !, question mark ?, parentheses (), and «» for quotation marks, although "" is also common. Hyphens (-) and the ellipsis (...) are also commonly found in modern writing.

Indication of tones [edit | edit source]

Lao

ພາສາລາວ *phasa lao*

ລາວ
(ພາສາລາວ)

Pronunciation	pʰáːsǎː láːw
Native to	Laos, Isan
Ethnicity	Lao Isan
Native speakers	30 million[1]
Language family	Kra–Dai • Tai • Southwestern Tai (Thai) • Lao–Phutai • **Lao**
Writing system	Lao script in Laos Thai script in Thailand Thai and Lao Braille
Official status	
Official language in	Laos ASEAN[2]
Recognised minority language in	Thailand (Isan) Cambodia (native to provinces of Steung Treng, Preah Vihear, Ratanakiri)
Language codes	
ISO 639-1	lo
ISO 639-2	lao
ISO 639-3	Either: lao – Laotian Lao tts – Isan (Thailand Lao)
Glottolog	laoo1244 Lao[3] nort2741 Northeastern Thai[4]
Linguasphere	47-AAA-c

Wikipedie je hyperpolyglot.

Hovoří 307 jazyky…

1 000 000+ articles

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users
1	English	English	en	**5,973,303**	49,054,097	922,807,885	1,150	37,675,749	138,(
2	Cebuano	Sinugboanong Binisaya	ceb	**5,378,800**	9,743,943	29,550,111	6	62,865	
3	Swedish	Svenska	sv	**3,744,904**	7,764,733	46,497,692	59	699,987	2,5
4	German	Deutsch	de	**2,367,768**	6,618,034	192,929,387	190	3,326,433	19,3
5	French	Français	fr	**2,157,680**	10,462,265	164,296,748	164	3,617,220	20,2
6	Dutch	Nederlands	nl	**1,983,360**	4,142,861	54,961,432	38	1,033,038	4,(
7	Russian	Русский	ru	**1,580,133**	6,039,378	103,179,135	89	2,636,165	11,3
8	Italian	Italiano	it	**1,566,626**	6,445,487	108,738,208	106	1,899,339	8,7
303	Sichuan Yi	ꆈꌠꁱꂷ	ii	**3**	188	11,652	1	1,586	1	0	
304	Afar	Afar	aa	**1**	509	4,680	1	3,328	1	0	
305	Muscogee	Muskogee	mus	**1**	114	3,600	1	1,682	1	0	

0 articles in language (other than main page)

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users	Files	De
306	Herero	Otsiherero	hz	**0**	175	4,480	1	3,121	1	0	
307	Kanuri	Kanuri	kr	**0**	161	4,640	1	4,475	1	0	

Proč dávat Wikipedii do korpusu?

Jaké jsou její výhody?

Obsahuje velké množství textu,

přes 50 milionů článků

kvalitního textu,

gramatika, pravopis, styl

v mnoha
různých jazycích,

včetně exotických

psaného množstvím
různých lidí,
má reprezentativní formu,
byť jen jediný žánr

pokrývajícího
všemožná témata,

má reprezentativní obsah,
byť ne ideálně vyvážený

sémanticky klasifikovaného,

kategorizace, Wikidata

souvislostně propojeného,

vnitřní odkazy, přesměrování

mezijazykově provázaného,

mezijazykové odkazy

jednotně formátovaného,

vzhled a styl, wikitext

strojově snadno zpracovatelného,

dumps, Parsoid/API, HTML5 DOM

licenčně bezproblémového,

CC BY-SA

známého původu a široce dostupného.

Wikimedia

Wikipedie do každého korpusu!

Už se tak děje, vkládá se na začátek.

Wiki2corpus

- Počítačový skript (Python, MIT licence)
  - `http://corpus.tools`
- Stáhne vybranou verzi Wikipedie a vyrobí korpus (prevertikál)
- Navazující nástroje:
  - čištění, tokenizace, deduplikace,
  - značkování, korpus (Sketch Engine)
- Nová verze 2.0:
  - využívá MediaWiki Parsoid API

Korpus české Wikipedie

22. listopadu 2019

COUNTS

Tokens	204,768,845
words	153,245,627
Sentences	23,800,279
Paragraphs	18,676,429
Documents	439,078

Nejčastější slova na české Wikipedii

1	v	5,316,874	18	pro	534,344	35	článek	278,423
2	a	4,245,817	19	ten	508,469	36	také	277,589
3	být	3,875,034	20	po	496,907	37	externí	274,420
4	na	2,393,033	21	mít	455,686	38	či	273,596
5	se	2,216,190	22	odkaz	440,235	39	reference	271,220
6	z	1,404,121	23	svůj	436,589	40	až	264,277
7	rok	1,362,110	24	za	432,366	41	dva	263,242
8	s	1,333,909	25	český	383,559	42	ale	259,876
9	který	981,978	26	jeho	362,808	43	místo	234,783
10	do	923,566	27	že	349,687	44	další	218,997
11	k	794,897	28	léto	345,462	45	část	218,067
12	on	641,053	29	stát	337,844	46	jeden	216,546
13	o	600,922	30	první	302,980	47	obec	214,424
14	jako	594,244	31	nebo	298,626	48	moct	213,569
15	i	578,394	32	u	292,064	49	mezi	201,720
16	tento	567,706	33	Praha	289,533	50	při	200,844
17	od	552,630	34	město	287,533			

Nejčastější oproti obecnému jazyku

KEYWORDS [DEV] Wiki2Corpus 2.0 C

SINGLE-WORDS ✓ MULTI-WORDS ✓

reference corpus: Czech Web 2017 (csTenTen17)

	Word		Word		Word		Word		Word
1	Commonsa	11	Division	21	souborný	31	Wikicitát	41	GBR
2	Wikimedium	12	Databasa	22	antuka	32	hebrejsky	42	km2
3	displaystyl	13	VG	23	FRA	33	ITA	43	územně
4	reference	14	departement	24	Movie	34	Ottův	44	dvouhra
5	ISBN	15	Wikizdroj	25	NGC	35	URS	45	ZOH
6	externí	16	úč	26	Libri	36	WTA	46	mathbf
7	Wikislovník	17	podbarvení	27	biografie	37	kanton	47	ČÚZK
8	LOH	18	slovníkový	28	mathrm	38	C	48	K
9	Wikipedie	19	anglicky	29	Football	39	JPN	49	M
10	frac	20	OG	30	vyd	40	Wikidruze	50	Gera

Rows per page: 50 1–50 of 1,000 1

Jaké všeliké sekce „Odkazy“ máme

	Word	Frequency	Frequency per million
1	Externí odkazy	270,472	1,320.86
2	Související odkazy	438	2.14
3	Kulturní odkazy	71	0.35
4	Další odkazy	36	0.18
5	Podobné odkazy	22	0.11
6	externí odkazy	21	0.10
7	Vnější odkazy	11	0.05
8	Externé odkazy	10	0.05
9	Externi odkazy	9	0.04
10	České odkazy	6	0.03
11	Extrení odkazy	6	0.03
12	Extermí odkazy	5	0.02
13	Cizojazyčné odkazy	5	0.02
14	Jiné odkazy	4	0.02
15	Exsterní odkazy	4	0.02
16	Internetové odkazy	3	0.01
17	Extrerní odkazy	3	0.01
18	Externím odkazy	3	0.01
19	Externaí odkazy	3	0.01
20	Exeterní odkazy	3	0.01

25	Reference odkazy	2	< 0.01	•••
26	Ostatní odkazy	2	< 0.01	•••
27	Multimediální odkazy	2	< 0.01	•••
28	Mediální odkazy	2	< 0.01	•••
29	Interní odkazy	2	< 0.01	•••
30	Geografické odkazy	2	< 0.01	•••
31	Extérmní odkazy	2	< 0.01	•••
32	Exterrní odkazy	2	< 0.01	•••
33	Externe odkazy	2	< 0.01	•••
34	Esterní odkazy	2	< 0.01	•••
35	Ecterní odkazy	2	< 0.01	•••
36	Audio odkazy	2	< 0.01	•••
37	Zahraniční odkazy	1	< 0.01	•••
38	Týmové odkazy	1	< 0.01	•••
39	Studijní odkazy	1	< 0.01	•••
40	Starověké odkazy	1	< 0.01	•••

45	Příbuzné odkazy	1	< 0.01	•••
46	Pozdější odkazy	1	< 0.01	•••
47	Poválečné odkazy	1	< 0.01	•••
48	Politické odkazy	1	< 0.01	•••
49	PExterní odkazy	1	< 0.01	•••
50	Oficiální odkazy	1	< 0.01	•••
51	Literární odkazy	1	< 0.01	•••
52	Kritické odkazy	1	< 0.01	•••
53	Jednotlivé odkazy	1	< 0.01	•••
54	Exzerní odkazy	1	< 0.01	•••
55	Extwrní odkazy	1	< 0.01	•••
56	Extetrní odkazy	1	< 0.01	•••
57	Extetní odkazy	1	< 0.01	•••
58	Extertní odkazy	1	< 0.01	•••
59	Externíí odkazy	1	< 0.01	•••
60	Externní odkazy	1	< 0.01	•••

	Word	↓ Frequency	Frequency per million
61	Extermní odkazy	1	< 0.01
62	Exterení odkazy	1	< 0.01
63	Exterbní odkazy	1	< 0.01
64	Extarní odkazy	1	< 0.01
65	Evterní odkazy	1	< 0.01
66	Etérmní odkazy	1	< 0.01
67	Eterní odkazy	1	< 0.01
68	Estení odkazy	1	< 0.01
69	Enterní odkazy	1	< 0.01
70	Doplňující odkazy	1	< 0.01
71	Biografické odkazy	1	< 0.01
72	Bibliografické odkazy	1	< 0.01
73	Biblické odkazy	1	< 0.01

Nejzdrojovanější slova

	Lemma	Frequency	Frequency per million
1	zdroj	520	38.23
2	obyvatel	220	16.18
3	léto	214	15.73
4	fotograf	189	13.90
5	USA	182	13.38
6	Praha	152	11.18
7	čtyřhra	110	8.09
8	místo	108	7.94
9	strana	107	7.87
10	rok	107	7.87
11	republika	107	7.87
12	zem	106	7.79
13	století	88	6.47
14	člověk	85	6.25
15	l.	85	6.25
16	km2	83	6.10
17	hráč	83	6.10
18	svět	81	5.96
19	m	75	5.51
20	Angeles	69	5.07

Nejčastější přídavná jména

WORDLIST [DEV] Wiki2Corpus 2.0 Cz

adjective (15,294 items | 1,211,110 total frequency)

	Lemma	Frequency		Lemma	Frequency
1	český	28,876	11	velký	8,380
2	americký	27,193	12	československý	8,018
3	externí	17,390	13	vysoký	8,013
4	německý	14,958	14	britský	7,903
5	francouzský	14,599	15	hlavní	7,751
6	anglický	11,482	16	slovenský	7,723
7	dobrý	11,134	17	ruský	7,581
8	další	11,091	18	italský	7,262
9	hudební	10,378	19	národní	6,861
10	nový	9,083	20	rakouský	6,844

Rows per page: 20 1–20 of 15,294 1

Typická slovní spojení k národnostem

WORD SKETCH

[DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

český as adjective 383,559×

modifiers of "český"

nouns modified by "český"

republika
České republiky

Budějovice
České Budějovice

zem
v českých zemích

země
země česká , kraj

televize
České televize

lípa
Česká Lípa

"český" and/or ...

československý
byl český a československý

slovenský
českých a slovenských

rakouský
byl rakouský a český

moravský
českých a moravských

německý
českou a německou

uherský
český a uherský

... is "český"

titul
Obhájcem titulu byl český pár

Praha
Praha byl český

autor
autorem je český

otec
Jeho otcem byl český

manžel
Jejím manželem byl český

manželka
Jeho manželkou byla česká

words before "český"

také
také český

i
i český

respektive
respektive České

jako
jako české

též
též Česká

zejména
zejména české

WORD SKETCH

[DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

americký as adjective 132,344×

modifiers of "americký"

nouns modified by "americký"

stát
Spojené státy americké

herec
americký herec

herečka
americká herečka

film
americký film

zpěvák
americký zpěvák

spisovatel
americký spisovatel

"americký" and/or ...

britský
americké a britské

evropský
evropských a amerických

kanadský
amerických a kanadských

anglický
anglické a americké

francouzský
francouzské a americké

sovětský
americké a sovětské

... is "americký"

titul
dvojic . Obhájcem titulu byl americký

Corporation
Corporation je americká

Compana
Company je americká

autor
autorem je americký

Recordsa
Records je americká hudební nahrávací společnost

otec

words before "americký"

zejména
zejména amerických

též
též americký

také
také americký

hlavně
hlavně americké

rovněž
rovněž americký

i
i americký

WORD SKETCH

[DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

ruský as adjective 48,446×

modifiers of "ruský"

nouns modified by "ruský"

federace
Ruské federace

impérium
Ruské impérium

car
ruský car

Wikipedie
na ruské Wikipedii . Externí odkazy

spisovatel
ruský spisovatel

armáda

"ruský" and/or ...

sovětský
sovětský a ruský

ukrajinský
ruských a ukrajinských

polský
polské a ruské

německý
německé a ruské

čínský
velká encyklopedie ruského a čínského letectví

italský

words before "ruský"

zejména
zejména ruských

především
především ruských

také
také ruský

hlavně
ruské literatury starali hlavně ruští básníci a méně

i
i ruský

dokonce

... is "ruský"

titul
hráček . Obhájkyní titulu byla ruská tenistka

stanice
Velitelem stanice byl ruský kosmonaut Jurij

WORD SKETCH

[DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

francouzský as adjective 74,972×

modifiers of "francouzský"

nouns modified by "francouzský"

kanton
je francouzský kanton v departementu

Wikipedie
na francouzské Wikipedii . Externí odkazy

král
francouzského krále

obec
je francouzská obec v departementu

revoluce
Velké francouzské revoluce

"francouzský" and/or ...

britský
britské a francouzské

anglický
anglických a francouzských

italský
francouzských a italských

německý
francouzské a německé

španělský
francouzské a španělské

navarrský
francouzská a navarrská

... is "francouzský"

titul
dvojic . Obhájcem titulu byl francouzský pár

autor
autorem je francouzský

otec
otcem byl francouzský

words before "francouzský"

zejména
zejména francouzské

také
také francouzský

co
co francouzský

i
i francouzský

především
především francouzské

hlavně
hlavně francouzští

WORD SKETCH

[DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

vlámský as adjective 1,462×

modifiers of "vlámský"

nouns modified by "vlámský"

Brabant
Vlámský Brabant

kartograf
vlámský kartograf

liberál
Vlámští liberálové a demokraté

malíř
vlámský malíř

aliance
Nová vlámská aliance

region
ve Vlámském regionu

"vlámský" and/or ...

křesťanskodemokratický
Křesťanskodemokratická a vlámská strana (Christen-Democratisch

holandský
holandských a vlámských umělců

nizozemský
nizozemských a vlámských umělců

valonský
Vlámský a Valonský

frankofonní
Bruselského regionu na frankofonní a vlámskou část se zákazem

normanský

words before "vlámský"

rovněž
tři společenství a rovněž vlámský i valonský region

také
také vlámský

Kolik a kterých článků jsem editoval (Blahma)

	Title	Frequency	Relative [%]
41	Adresář (seznam)	1	12,895.7
42	Adrian Dietrich Lothar von Trotha	1	5,170.9
43	Adrian von Arburg	1	2,214.6
44	Aelbert Cuyp	1	6,640.2
45	Aernout van der Neer	1	5,409
46	Aeronet.cz	1	4,163.2
47	Afrika (divadelní hra)	1	12,367.8
48	Agent Státní bezpečnosti	1	41,468.6
49	Airbag	1	5,572.9
50	Aisne (přítok Oise)	1	38,452.7
51	Ajznbonský tovaryš	1	6,800.3
52	Akademický podvod	1	36,152.1
53	Akademický senát Masarykovy univerzity	1	12,189.6
54	Akademie esperanta	1	15,724.2
55	Akce D.O.S.T.	1	4,401.5
56	Akcie na majitele	1	3,831.3
57	Akt bezpodmínečné kapitulace nacistického Německa	1	14,146.5
58	Akta X	1	6,497.4
59	Albert Claude	1	136,445.1
60	Albánská vlajka	1	9,701.4

Typická slova z článků, které jsem editoval

reference corpus: [DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

	Word		Word		Word		Word		Word
1	esperanto	11	Babiš	21	KSM	31	Haploskupin	41	pochválen
2	esperantský	12	Toufar	22	migrant	32	Valách	42	Řečkovice
3	esperantista	13	MU	23	Veligrad	33	velehradský	43	Bystrc
4	dront	14	lachema	24	SFŘ	34	Titanicus	44	Buchlovice
5	Hlučínsko	15	videoarchiv	25	Kofola	35	Bolatice	45	Vaccinium
6	ostrožský	16	Přibyslav	26	bělka	36	pokémon	46	ACTA
7	vrut	17	SANEP	27	Toufarův	37	Chřiby	47	Zdechovský
8	Zamenhof	18	Antonínek	28	Tisa	38	Číhošť	48	orloj
9	kuratorium	19	Baltík	29	Laudon	39	máz	49	Trumpův
10	ICQ	20	Velehrad	30	Špilberk	40	Trump	50	Medlánka

Rows per page: 50 1–50 of 1,000 1

Typická slova z článků, které editovala KKDAII

reference corpus: [DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

	Word		Word		Word		Word		Word
1	přípora	11	Fugger	21	ostění	31	sbíhat	41	předsíň
2	kružba	12	výžlabek	22	Coventr	32	hrotitý	42	plaménkový
3	svorník	13	trojlist	23	sedile	33	žebr	43	zaklenut
4	vegetabilní	14	presbytář	24	pětiboký	34	Rejt	44	čtyřlist
5	lomený	15	trojlodí	25	sakristie	35	patka	45	ambit
6	Etna	16	zaklenout	26	žebrový	36	konzola	46	trojloď
7	opěrák	17	manuelský	27	vyžlabený	37	Dauchra	47	kroužený
8	presbyterium	18	klenba	28	fiála	38	Palácio	48	Líbal
9	žebro	19	zaklenutí	29	litomyšlský	39	polygonální	49	síňový
10	klenební	20	kruchta	30	profilovaný	40	obloun	50	podokenní

Rows per page: 50 1–50 of 1,000 1

Typická slova z článků, které editoval Frettie

reference corpus: [DEV] Wiki2Corpus 2.0 Czech (cswiki2corpus)

	Word		Word		Word		Word		Word
1	třebíčský	11	borovina	21	TKO	31	participated	41	Křižanovský
2	videoarchiv	12	rozpoznaný	22	P3	32	Semir	42	Lubnice
3	Třebíč	13	Dědice	23	třeštit	33	Dačice	43	CSSD
4	Vantuch	14	Dukovany	24	LSSAH	34	Bíteš	44	UFC
5	Matton	15	dns	25	podklášteří	35	Lancra	45	ČT
6	Jemnice	16	tele	26	horácký	36	Třebíčsko	46	brtnický
7	Rokytný	17	Jaroměřice	27	Rouchovan	37	Okříšek	47	Hobzí
8	Brtnice	18	Želetava	28	Telč	38	Zones	48	events
9	Hrotovice	19	LV	29	FIBA	39	NDMS	49	auth
10	NBL	20	Náměšť	30	úhrnem	40	P2	50	Tasov

Rows per page: 50 1–50 of 1,000 1

A kolik je hrabat?

Sloveso: 8 článků

- Dinosauři
- Fosilie
- Kur domácí
- Sysel Parryův
- Tragédie na Mt. Hood
- Zajíc polární
- + 2 filmy

Podstatné jméno:
920 článků

- Rody
- Státy
- Dějiny
- Heraldika
- Místopis
- Hrady
- Muzea…

https://commons.wikimedia.org/wiki/File:PEAKE(1874)_p387_Leopold_I..jpg – Elizabeth Peake, PD

www.sketchengine.eu

Děkuji za pozornost!

Marek Blahuš
<wikipedia@blahus.cz>