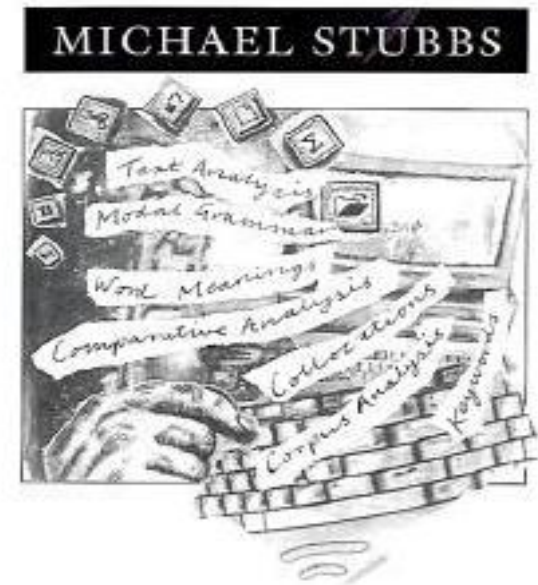


ABC o jazykovém korpusu: jak pracovat se SketchEngine



TEXT AND
CORPUS ANALYSIS

Co je jazykový korpus?

Korpus – elektronický soubor textů (mluvených nebo psaných), v němž je možné vyhledávat jazykové jevy

- ***Korpusy dle velikosti***
 - velké korpusy obecného jazyka (milióny slov)
 - malé specializované korpusy (do 250 000 slov)
- ***Anotace korpusu*** – manuální nebo automatická
 - ***tagging*** – připojují se morfologické značky (tagy), např. slovní druhy a gramatické kategorie slov
 - ***lematizace*** – přidělování každého slovního tvaru v korpusu k lemmatu, tj. abstrakce slova s jeho morfologickými vlastnostmi
 - ***parsing*** – vytváří se stromové databanky umožňující rozeznání syntaktických funkcí slov ve větné stavbě

Příklady obecních korpusu

BNC – Britský národní korpus (1990 –)

- 100 milionů slov (90% psané a 10% mluvené texty)
- literatura, noviny, vědecké, právnické atd. texty

COCA – Korpus současné americké angličtiny (1990 – 2017)

- 560 milionů slov
- rovnoměrně reprezentovaný mluvený jazyk, literatura, časopisy, noviny a odborné texty

ČNK – Český národní korpus (2000 –)

- 4 mld. slov (80% psané a 20% mluvené texty)
- diachronní korpus starších textů
- InterCorp – paralelní korpus překladů

SketchEngine

SketchEngine je komerční korpusový nástroj pro lingvisty, učitele, lexikografy a překladatele, který umožňuje hledat ve více než 80 jazycích.

- Přístup k velkému množství korpusů zdarma
- Možnost rychle najít příklady z autentického jazyka v kontextu
- Vyhledat gramatické chování a význam slov a frází
- Možnost vytvořit vlastní korpus z textů ve formátech .doc, .rtf, .pdf, .txt a z internetových zdrojů
- Automatická anotace korpusu



Přihlášení k SketchEngine

<https://ske.fi.muni.cz>

Sketch Engine

Home
Log in
Lost password

Support

User guide
Contact support

Log in

Masaryk University staff and students

SSO authentication ?

External account holders

User name

Password

Log in

[Password recovery](#)

Přihlášení k SketchEngine

ENGLISH 



MASARYKOVA
UNIVERZITA

Jednotné přihlášení na MUNI

ske.fi.muni.cz

 **UČO / GUEST ID**

 **SEKUNDÁRNÍ HESLO**

PŘIHLÁSIT

Tato služba vyžaduje ověření Vaší identity
(**UČO / Guest ID**)

Pokud neznáte **sekundární heslo**, můžete si je
nastavit prostřednictvím IS MU na stránce [změna
hesla](#).

V případě problémů či dotazů kontaktujte prosím helpdesk@ics.muni.cz.

Korpusy v SketchEngine

Home

+ Create corpus

+ WebBootCaT

+ Upload TMX or XLS

Parallel corpora

Compare corpora

My jobs

Advanced features

Corpus templates

Sketch grammars

Subcorpus definitions

User groups

Subscription overview

Support

User guide

Feedback

Corpora:

Recent

My own

Shared with me

Featured

Parallel

All

Search:




Filter by language: all

Language	Name	Words		
English	123	61,961		
English	Academia PL-EN Parallel Corpus (English)	358,478		
Polish	Academia PL-EN Parallel Corpus (Polish)	293,654		
English	ACL Anthology Reference Corpus (ARC)	38,795,751		
-- other (UTF-8) --	Amharic WaC [2013 + 2015]	15,217,564		
-- other (UTF-8) --	Amharic WaC [2013]	8,772,463		
Arabic	Arabic web corpus	407,005		
English	Araneum Anglicum Maius (En Web 14.04) (with LCL sketches)	769,606,166		
English	Araneum Anglicum Maius (En Web 14.04) 1,20 G	769,606,166		
English	Araneum Anglicum Minus (En Web 14.04) 119 M	91,938,147		
Czech	Araneum Bohemicum III Maius (Czech, 17.04) 1,20 G	998,957,244		
Czech	Araneum Bohemicum III Minus (Czech, 17.04) 120 M	99,834,283		
French	Araneum Francogallicum Maius (Fr Web 14.04) 1,23 G	777,825,756		
French	Araneum Francogallicum Minus (Fr Web 14.04) 125 M	95,941,088		
German	Araneum Germanicum Maius (De Web 14.04) 1,20 G	711,223,591		
German	Araneum Germanicum Minus (De Web 14.04) 120 M	88,028,254		
Dutch	Araneum Nederlandicum Maius (NL Web 14.04) 1,20 G	713,417,518		
Dutch	Araneum Nederlandicum Minus (NL Web 14.04) 120 M	90,887,838		
Polish	Araneum Polonicum Maius (PL Web 14.04) 1,04 G	595,768,667		
Polish	Araneum Polonicum Minus (PL Web 14.04) 110 M	78,493,398		
Russian	Araneum Russicum Maius (Ru Web 14.04) 1,22 G	767,777,339		
Russian	Araneum Russicum Minus (Ru Web 14.04) 123 M	89,041,788		
Slovak	Araneum Slovaccum Maius (Sk Web 14.04) 1,20 G	521,850,341		
Slovak	Araneum Slovaccum Minus (Sk Web 14.04) 121 M	71,370,122		
English	Barack Obama	738,094		
Bengali	BengaliWaC	303,095		
Polish	BIBLE Polish, plahili-Polish	138,216		
Swahili	BIBLE Swahili, Swahili-Polish	139,160		
English	British Academic Spoken English Corpus (BASE)	1,186,290		
English	British Academic Written English Corpus (BAWE)	6,968,089		
English	British National Corpus	96,048,950		
English	British National Corpus (BNC)	96,133,793		
English	British National Corpus (TreeTagger, xkovar3 testing)	96,267,546		

Co pro nás SketchEngine může udělat?

- Frekvenční seznam slov
- Tezaury
- Slovní profily (*word sketches*)
- Konkordance – výskyty daného jevu s okolním kontextem
 - gramatické vazby – *I think that, I am thinking of*
 - kontextualizace významu
white board, white wine, white coffee, white collar
 - kolokace
confess a sin, undergo surgery

Hledání v korpusu

 [British National Corpus](#)

[Home](#)
[Search](#)
[Word list](#)
[Word sketch](#)
[Thesaurus](#)
[Sketch diff](#)
[Trends](#)
[Corpus info](#)
[My jobs](#)
[User guide](#)

Simple query: [Make Concordance](#)

[Query types](#) [Context](#) [Text types](#) ?

Text types

Subcorpus: [info create new](#) ?

TEXT TYPE

- Spoken context-governed
- Spoken demographic
- Written books and periodicals
- Written miscellaneous
- Written-to-be-spoken

[Select All](#)

PUBLICATION DATE

- 1960-1974
- 1975-1984
- 1985-1993
- Unknown

[Select All](#)

DOMAIN FOR WRITTEN CORPUS TEXTS

- Imaginative
- Informative: applied science
- Informative: arts
- Informative: belief & thought
- Informative: commerce & finance
- Informative: leisure
- Informative: natural & pure science
- Informative: social science
- Informative: world affairs

[Select All](#)

MEDIUM FOR WRITTEN CORPUS TEXTS

- Book
- Miscellaneous: published
- Miscellaneous: unpublished
- Periodical
- To-be-spoken

[Select All](#)

Frekvenční seznam slov

Sketch Engine [British National Corpus](#) Sketch Engine [British Academic Written English Corpus \(BAWE\)](#)

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Trends
Corpus info
My jobs
User guide [↗](#)

Save
Change options

Menu position

Word list
Corpus: British National Corpus
Total number of items: 181,506
Page [Next >](#)

<u>word</u>	<u>Freq</u>
the	5,415,473
of	3,027,293
to	2,566,874
and	2,510,049
a	2,040,346
in	1,788,884
that	1,062,370
is	972,923
was	877,918
I	862,214
for	831,741
it	821,805
on	694,931
be	648,529
with	640,381
The	619,060
as	603,494
you	574,714
at	487,203
by	486,695
he	462,137
are	455,051
have	453,293
not	432,451
had	418,575
from	410,171
his	381,674
which	361,184
or	358,453
this	343,534
they	336,525

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Corpus info
My jobs
User guide [↗](#)

Save
Change options

Menu position

Word list
Corpus: British Academic Written English Corpus (BAWE)
Total number of items: 46,307
Page [Next >](#)

<u>word</u>	<u>Freq</u>
the	429,343
of	270,136
and	207,623
to	188,666
in	137,911
a	125,736
is	110,721
that	78,781
The	62,638
as	62,128
be	58,053
for	54,827
are	42,545
with	40,947
it	39,044
by	38,772
on	38,754
was	36,764
this	35,020
not	33,021
from	29,638
which	29,041
an	26,692
have	25,549
can	25,202
or	22,999
at	22,262
has	20,129
This	19,366
their	18,825
were	17,882
will	17,825

Tezaurus

Sketch Engine

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Trends
Corpus info
My jobs
User guide ↗

Clustering
Save

Menu position

obsolete

(*adjective*) Alternative PoS: [verb](#) (freq: 11) [noun](#) (freq: 2)
British National Corpus (BNC) freq = [330](#) (2.93 per million)

Lemma	Score	Freq
unusable	0.237	90
impotent	0.225	216
superfluous	0.212	236
inoperative	0.206	37
worthless	0.198	345
outdated	0.188	189
untenable	0.183	143
moribund	0.176	114
ineffective	0.176	664
ineffectual	0.175	171
unemployable	0.167	37
unenforceable	0.165	131
extinct	0.165	432
inaccessible	0.161	331
sterile	0.160	384
redundant	0.159	1,167
speechless	0.158	193
unprofitable	0.153	162
automated	0.150	308
comprehensible	0.150	203
usable	0.146	308
uninhabitable	0.144	37
meaningless	0.141	641
impracticable	0.139	167
collectable	0.135	55
indispensable	0.134	425
impassable	0.134	98
defunct	0.130	155
illegible	0.130	80
bearable	0.130	116
useless	0.124	1,264
unintelligible	0.123	133

A word cloud visualization of synonyms for the word 'obsolete'. The most prominent words are 'unusable', 'impotent', 'superfluous', 'inoperative', 'worthless', 'outdated', 'untenable', 'moribund', 'ineffective', 'ineffectual', 'unemployable', 'unenforceable', 'extinct', 'inaccessible', 'sterile', 'redundant', 'speechless', 'unprofitable', 'automated', 'comprehensible', 'usable', 'uninhabitable', 'meaningless', 'impracticable', 'collectable', 'indispensable', 'impassable', 'defunct', 'illegible', 'bearable', 'useless', and 'unintelligible'. The words are arranged in a circular pattern, with 'unusable' and 'impotent' being the largest.

Slovní profil

Sketch Engine

British National Corpus (BNC)

Home

Search

Word list

Word sketch

Thesaurus

Sketch diff

Trends

Corpus info

My jobs

User guide

Save

Change options

Cluster

Sort by freq

Hide gramrels

More data

Less data

Menu position

pay (verb) Alternative PoS: [noun](#) (4,665)
British National Corpus (BNC) freq = [38,062](#) (338.96 per million)

modifiers of "pay"	objects of "pay"	subjects of "pay"	"pay" and/or ...	prepositional phrases
5,168 0.14	17,176 0.45	5,295 0.14	465 0.01	7,052
low 58 8.44	attention + 1,400 10.82	customer 57 7.86	receive 24 9.61	"pay" for ... 3,041 0.08
low paid ,	tax + 687 10.13	consumer 43 7.83	received or paid	"pay" to ... 1,193 0.03
poorly 62 8.30	tribute + 543 9.94	consumers pay	book 10 9.33	"pay" by ... 735 0.02
poorly paid ,	pay tribute to	buyer 41 7.67	credit 7 8.88	"pay" in ... 644 0.02
dearly 25 7.23	price + 701 9.88	tenant 37 7.57	secure 6 8.36	"pay" on ... 342 0.01
paid dearly for	fee + 424 9.50	company + 133 7.54	transfer 6 8.20	"pay" at ... 177 0.00
highly 61 7.07	bill + 471 9.50	fee 35 7.53	buy 9 7.58	"pay" into ... 169 0.00
highly paid	rent + 375 9.39	fee paying	accept 6 7.47	"pay" under ... 106 0.00
actually + 122 7.04	visit + 369 9.25	worth 34 7.51	go 17 6.19	"pay" as ... 77 0.00
actually paid	dividend + 326 9.20	a price worth paying	go and pay	"pay" over ... 73 0.00
directly 46 6.96	money + 502 8.98	polluter 28 7.42	get 6 6.11	"pay" from ... 51 0.00
be paid directly to the	sum + 260 8.76	the 'polluter pays'	come 9 6.08	"pay" of ... 48 0.00
only + 304 6.90	rate + 308 8.64	employer 37 7.34	do 9 5.92	"pay" with ... 48 0.00
only pay	debt + 212 8.52	taxpayer 25 7.17		"pay" than ... 41 0.00
monthly 19 6.89	cost + 284 8.50	firm 44 7.06		"pay" through ... 40 0.00
paid monthly	pound + 228 8.50	firms pay		"pay" if ... 31 0.00
twice 27 6.85	wage + 203 8.47	tax 27 7.00		"pay" within ... 28 0.00
paid twice	interest + 287 8.40	tax paid		"pay" before ... 24 0.00
back + 182 6.80	pay interest	expense 22 6.98		"pay" towards ... 20 0.00
to pay back	contribution + 207 8.36	an all expenses paid		"pay" until ... 19 0.00
annually 19 6.74	premium + 173 8.31	member 50 6.95		"pay" per ... 18 0.00
paid annually	employment + 180 8.29	purchaser 22 6.94		"pay" during ... 17 0.00
already + 148 6.63	paid employment	paid by the purchaser		"pay" after ... 17 0.00
already paid	compensation + 166 8.23	amount 20 6.72		"pay" out ... 15 0.00
well + 162 6.56	to pay compensation	the amount paid		"pay" without ... 14 0.00

Konkordance

Sketch Engine

Home
Search
Word list
Word sketch
Thesaurus
Sketch diff
Trends
Corpus info
My jobs
User guide

Save
Make subcorpus
View options
KWIC
Sentence
Sort
Left
Right
Node
References
Shuffle
Sample

Query **pay** 42,727 (380.51 per million)

Page of 2,137 [Next](#) | [Last](#)

J2B	to be more work than charring - for less	pay	. Naturally she switched to being a char
J2B	proved to have prospects as poor as the	pay	. A change to teaching English to Italian
J2T	Guardian 8 February </p> Transport Dutch tolls to	pay	for transport <p> The Netherlands government
J2G	times during the past year. So many dollars	paid	to John Russell for so many stage horses
J2G	They heard him, you could tell, but didn't	pay	any attention. Of course not, I realized
J2G	be White or Mexican or Indian. But now it	pays	you to be a White man. To look like a White
J2G	It's not mine to sell." </p><p> " Then if I	paid	more than just our fares." </p><p> " You're
J2G	Mendez shrugged. " If they don't make it, who	pays	for them?" </p><p> " I buy the horses," Dr.
J2G	though I reminded him I was going too and he	paid	me my last wages. From then on I was no
J2G	here and say he's taking your seat-after	paying	your fare and all-and the company doesn't
J2G	walked out. </p><p> Braden didn't even offer to	pay	him for the ticket. He watched the ex-soldier
J2G	something to watch, I decided. To see if she	paid	any attention to him. </p><p> Frank Braden
J2C	, a contributory scheme to which members	pay	6.35 per cent of annual salary. Subject
J2C	to receive full-rate Statutory Maternity	Pay	or were at any stage entitled to the benefits
J2C	to the benefits of a previous employer's	paid	maternity leave scheme, women are eligible
J2C	eligible for 18 weeks maternity leave on full	pay	, followed by up to 22 weeks unpaid maternity
J2C	scheme is not linked to arrangements for	pay	, promotion or probation. Under the scheme
J2C	eligible for 18 weeks maternity leave on full	pay	, followed by up to 22 weeks unpaid maternity
J2C	of the successful candidate are generally	paid	in full in appropriate cases. Professional
J2C	not in charge of the department, shall be	paid	to the Curators of the University Chest

Page of 2,137 [Next](#) | [Last](#)

Porovnání slov



British National Corpus

- Home
- Search
- Word list
- Word sketch
- Thesaurus
- Sketch diff**
- Trends
- Corpus info
- My jobs
- User guide ↗

Change options

Menu position

handsome/beautiful (adjective)

British National Corpus freqs = 1,578 | 8,394

handsome 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 beautiful

and/or	609	2,017	0.39	0.24
debonair	3	0	7.2	--
well-built	3	0	7.1	--
dashing	3	0	7.0	--
arrogant	5	0	7.0	--
illustrated	3	0	6.7	--
smiling	3	0	6.4	--
athletic	3	0	6.3	--
lean	3	0	6.2	--
tall	36	12	7.7	5.8
dark-haired	4	4	7.2	5.9
blond	7	6	7.6	6.3
georgian	4	5	6.3	5.8
young	76	95	6.2	6.5
elegant	3	13	5.2	6.7
blonde	3	19	5.9	7.6
intelligent	3	17	5.2	7.1
peaceful	0	9	--	6.3
gracious	0	6	--	6.3
bright	0	21	--	6.3
beautiful	0	27	--	6.4
wild	0	20	--	6.5
sandy	0	8	--	6.5
historic	0	13	--	6.6
serene	0	7	--	6.7
sunny	0	11	--	6.8

modifier	326	1,869	0.21	0.22
devastatingly	5	0	8.7	--
darkly	6	0	8.5	--
ruggedly	3	0	8.2	--
outstandingly	3	4	7.9	6.1
strikingly	10	14	9.1	7.7
wonderfully	3	4	7.0	5.8
incredibly	4	10	7.1	7.0
yet	3	4	3.7	3.8
extraordinarily	4	14	7.3	7.6
rather	10	16	4.8	5.3
extremely	11	20	5.8	6.3
especially	3	9	4.7	5.8
very	73	274	4.9	6.7
as	13	85	3.9	6.5
so	27	189	4.0	6.8
most	29	533	4.3	8.5
absolutely	0	17	--	6.6
astonishingly	0	6	--	6.6
classically	0	6	--	6.6
startlingly	0	6	--	6.6
exceptionally	0	8	--	6.7
exquisitely	0	7	--	6.8
hauntingly	0	10	--	7.4
breathtakingly	0	15	--	8.0
stunningly	0	23	--	8.6

modifies	1,048	5,044	0.66	0.60
devil	10	0	6.7	--
miguelito	3	0	6.5	--
plumage	4	0	6.3	--
binder	3	0	6.0	--
prince	14	0	5.9	--
chap	7	0	5.9	--
admirer	3	0	5.7	--
dividend	7	0	5.5	--
fellow	7	0	5.5	--
singer	6	0	5.4	--
donation	4	0	5.3	--
mansion	4	7	5.5	5.1
face	87	75	5.6	5.3
woman	26	255	3.3	6.5
girl	7	123	2.6	6.6
flower	0	31	--	5.9
landscape	0	23	--	6.0
valley	0	30	--	6.0
bride	0	14	--	6.1
setting	0	26	--	6.1
beach	0	46	--	6.9
garden	0	104	--	6.9
surroundings	0	31	--	7.1
scenery	0	40	--	7.7
countryside	0	70	--	7.7

adj_subject	41	506	0.03	0.06
face	6	6	1.7	1.7
world	0	4	--	0.7

SketchEngine

<https://ske.fi.muni.cz>

