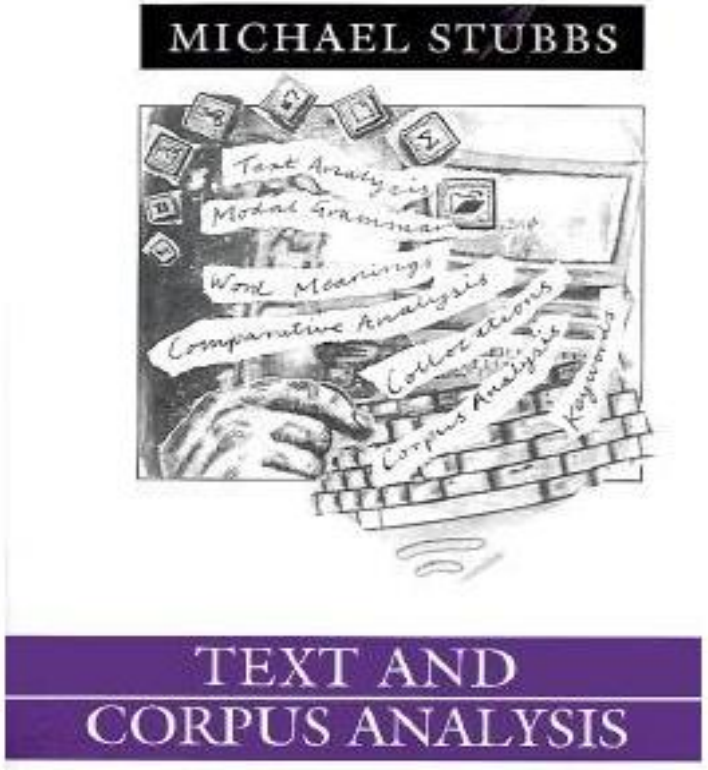
ABC o jazykovém korpusu: jak pracovat se SketchEngine



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 (miliony 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

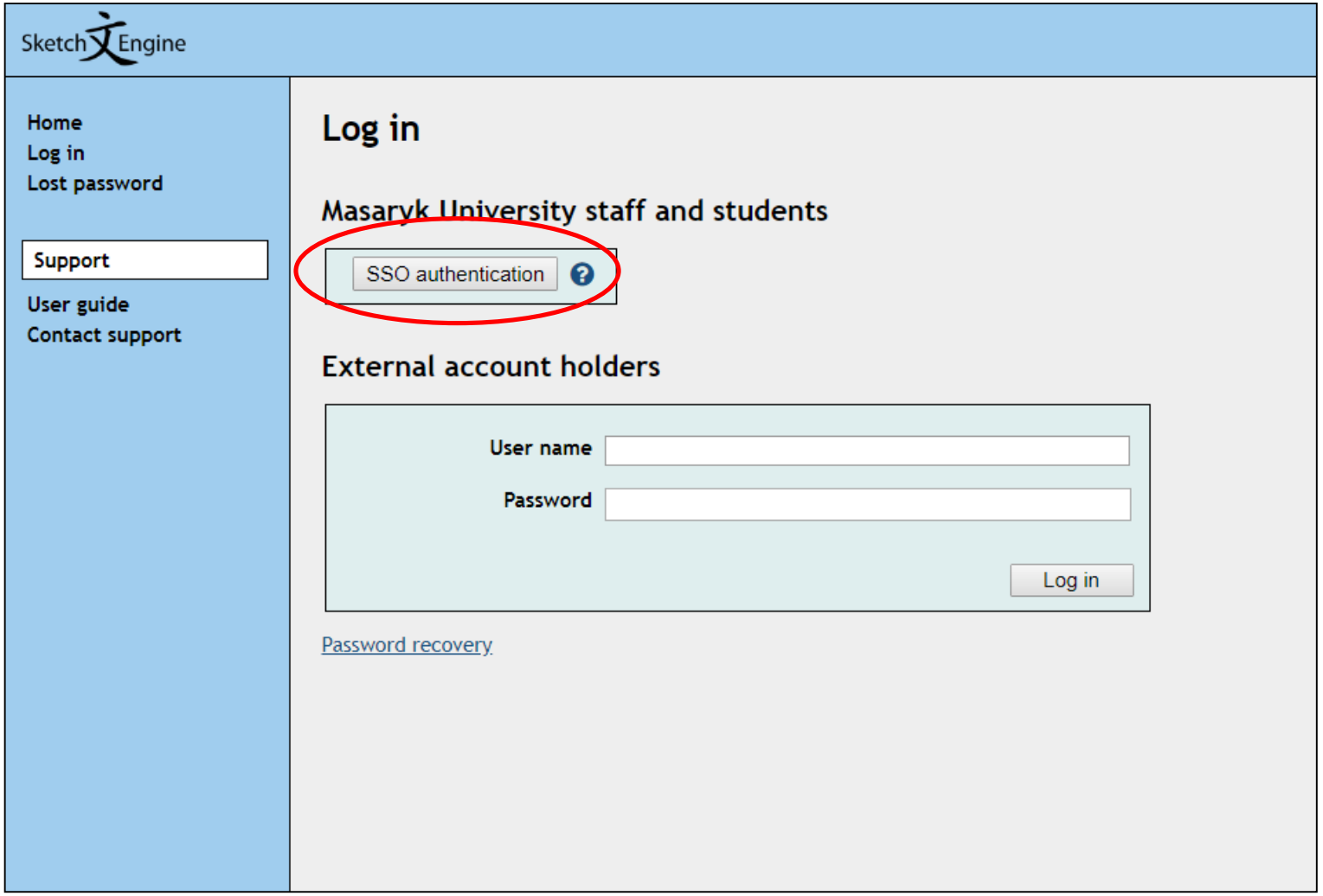
SketchEngine

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

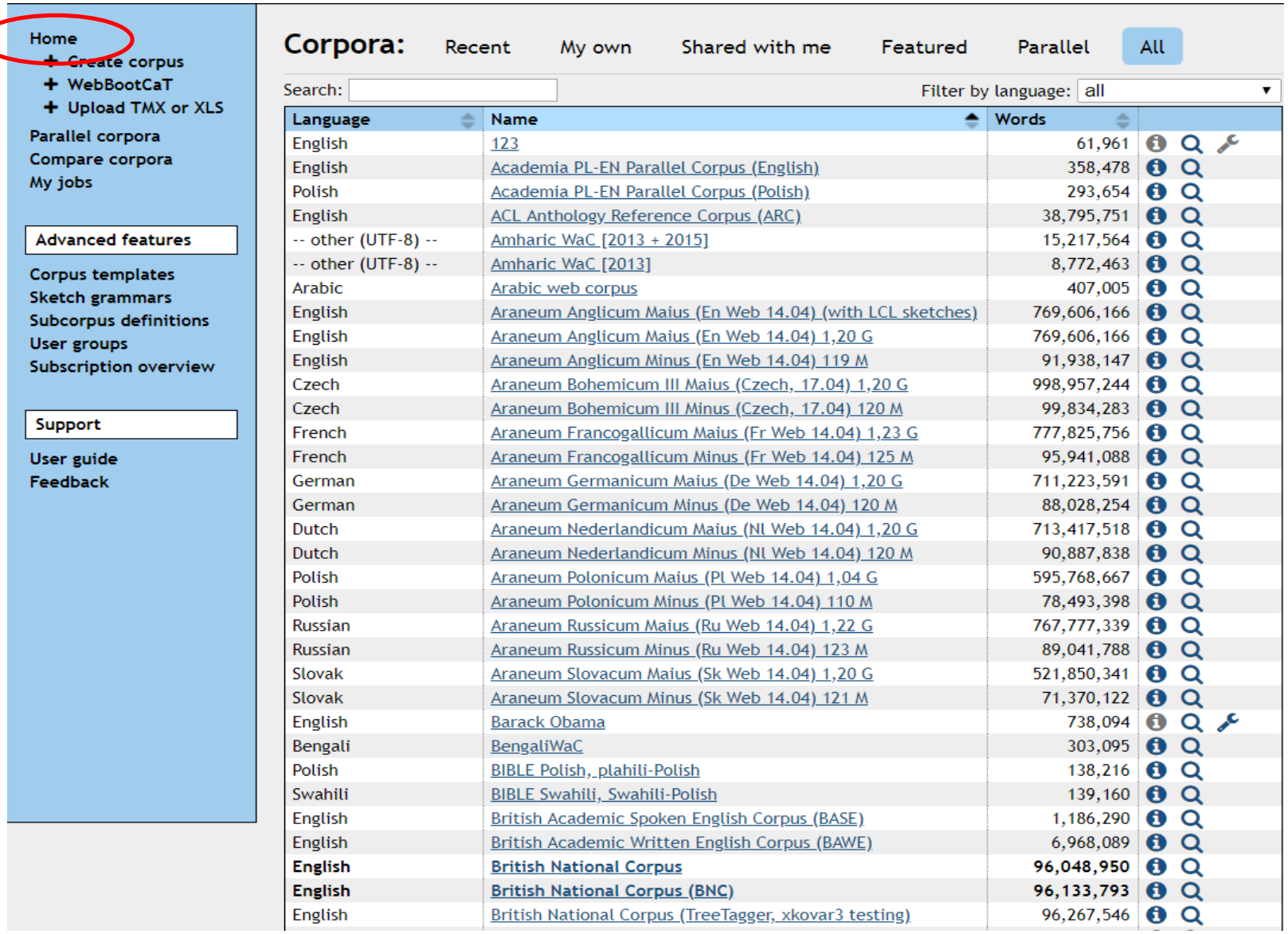
SKETCH ENGiNE

Přihlášení k SketchEngine
https://ske.fi.muni.cz



Přihlášení k SketchEngine

Korpusy v SketchEngine



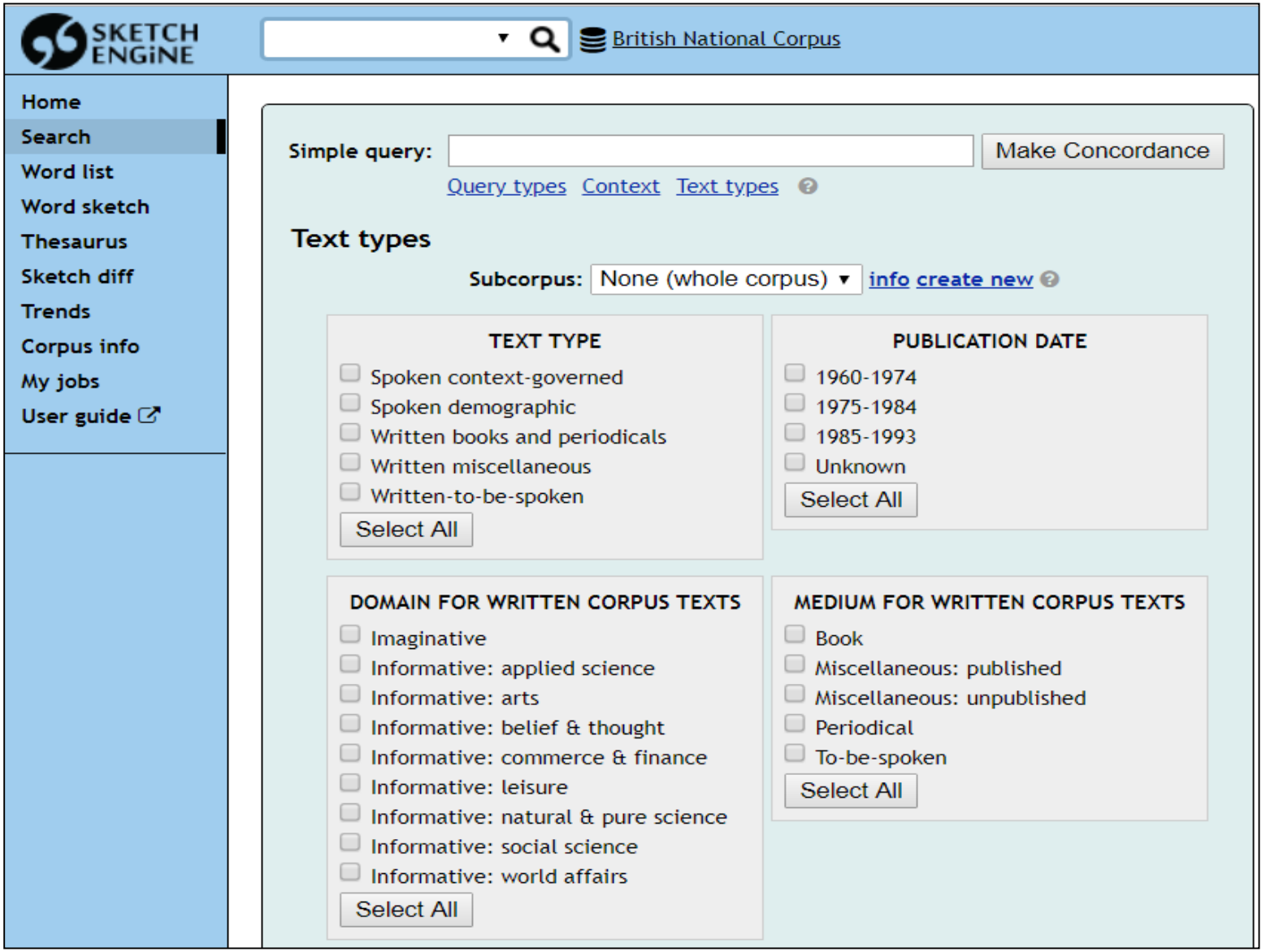
Co pro nás SketchEngine může udělat?

- 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

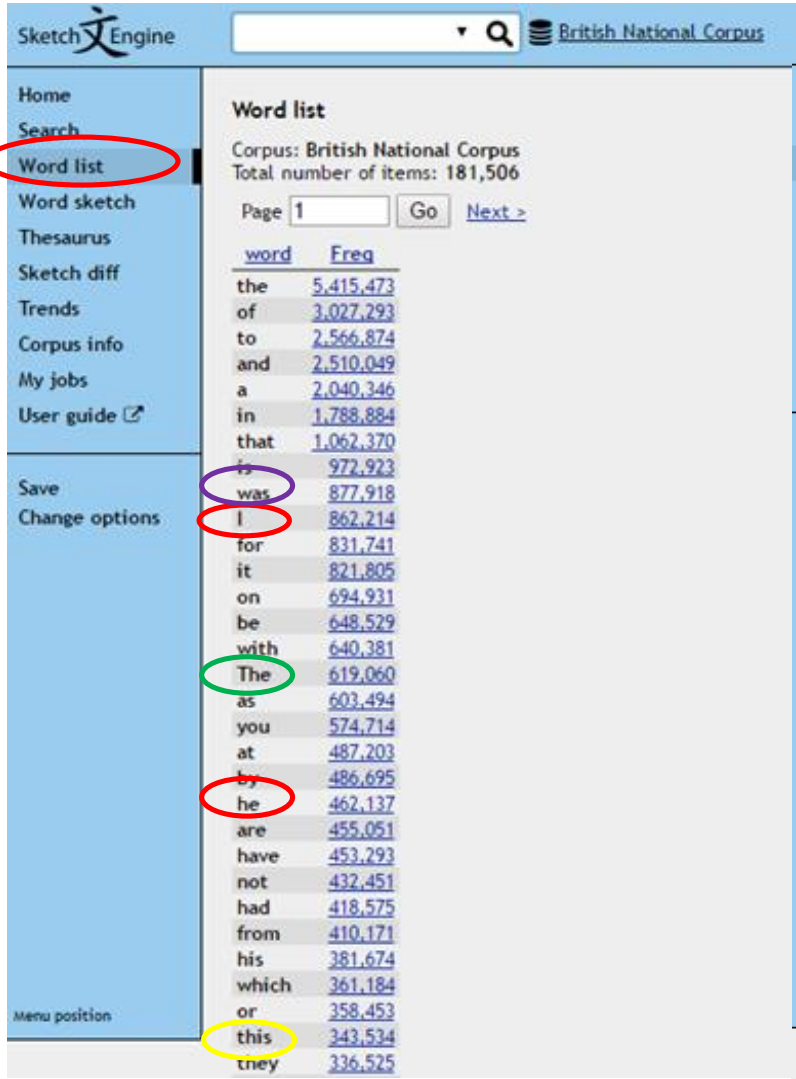
    white board, white wine, white coffee, white collar
  - kolokace

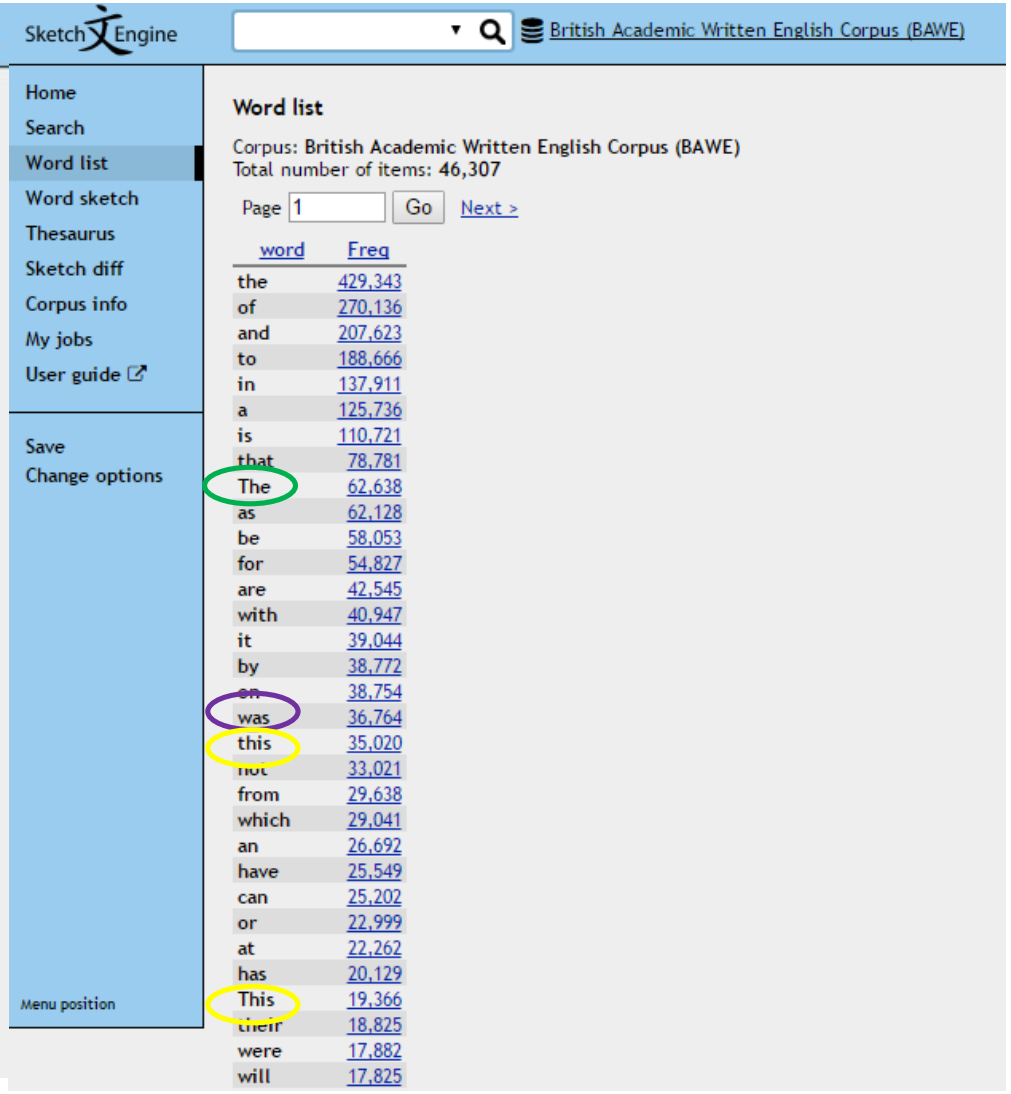
    confess a sin, undergo surgery

Hledání v korpusu



Frekvenční seznam slov





Tezaurus

Sketch Engine | British National Corpus (BNC)

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

- Clustering
- Save

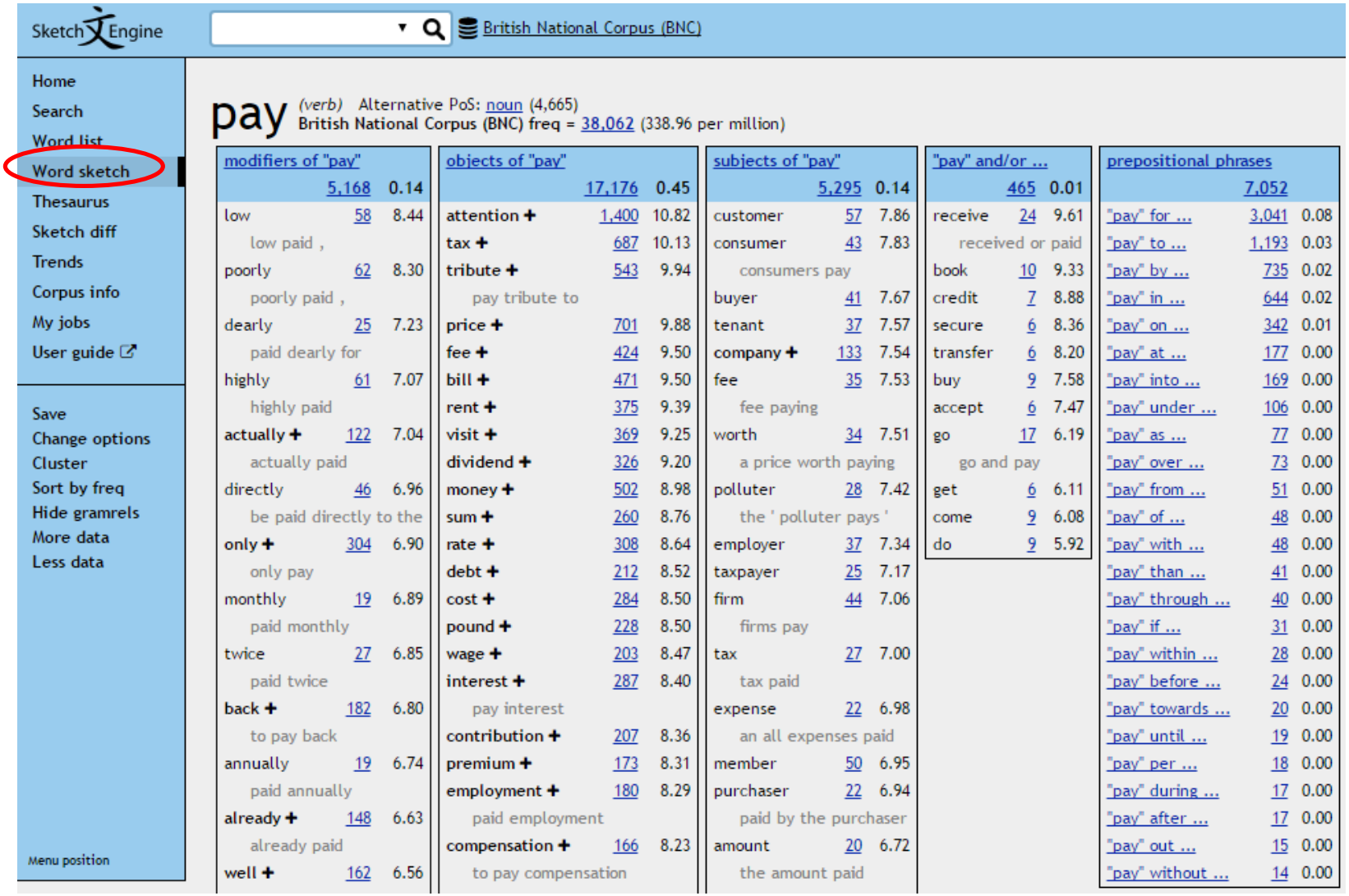
Menu position

obsolete *(adjective)* Alternative PoS: verb (freq: 11) noun (freq: 2)

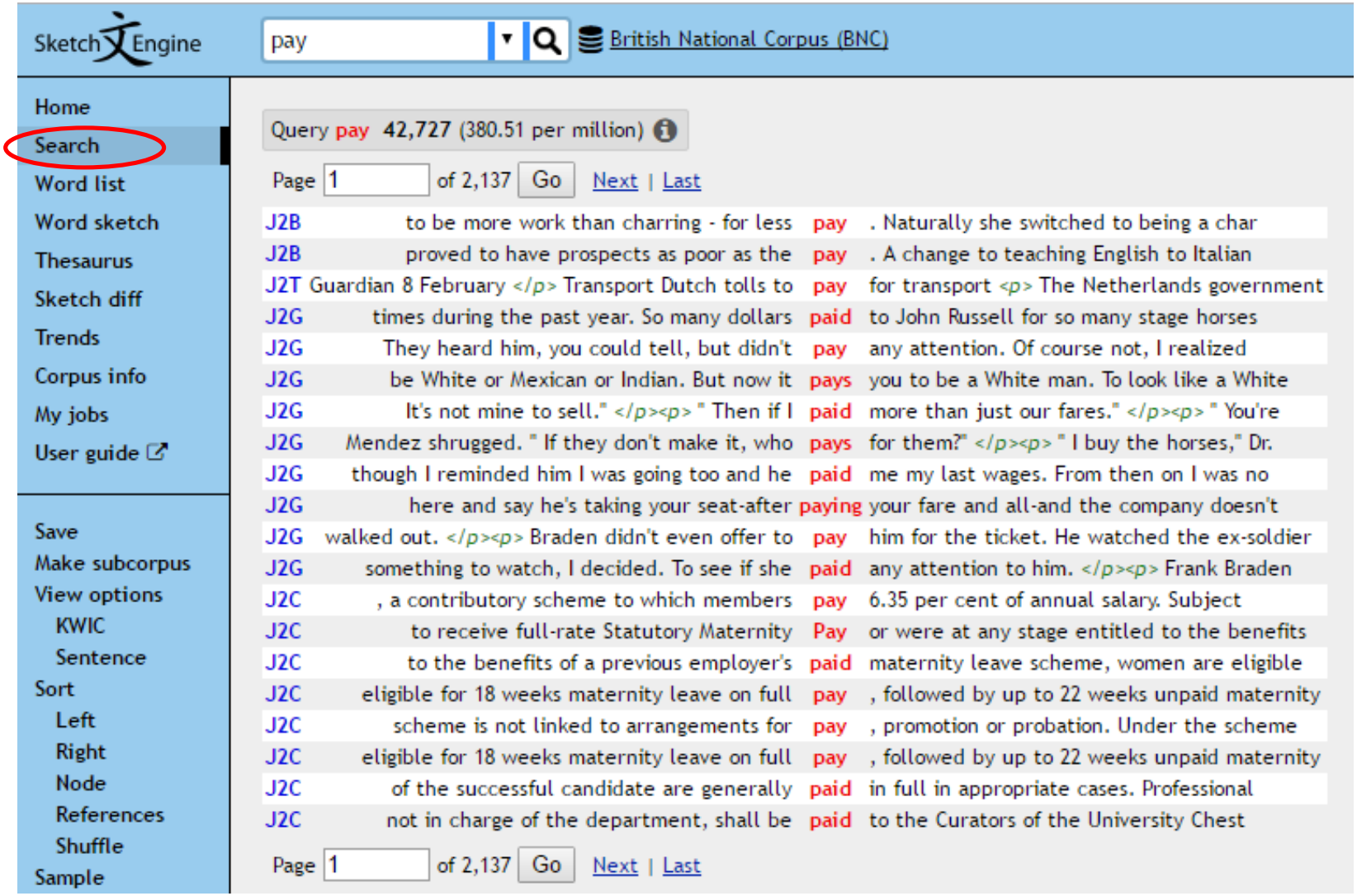
British National Corpus (BNC) freq = 330 (2.93 per million)

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

Slovní profil



Konkordance



Porovnání slov

SKETCH ENGINE

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

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

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

SketchEngine

https://ske.fi.muni.cz