Korpusová lingvistika – nástroje a jejich využití

Počítačové nástroje pro češtinu
Jaro 2020

Jakub Machura

415795@mail.muni.cz

Co nás čeká?

1. Nástroje běžící na pražské platformě **korpus.cz**
    - WaG
    - SyD
    - Morfio
    - Kwords
    - Treq

Co nás čeká?

2. Brněnský korpusový manažer **Sketch Engine**
    - Word Sketch
    - Tezaurus
    - N-gramy
    - Skell

WaG – Slovo v kostce

- aplikace poskytuje rychlý a základní přehled o tom, jak se v korpusu zadané slovo používá
- data získaná z psaného, mluveného i paralelního korpusu
- funkce: *Vyhledat slovo*

  Hledat ve dvou jazycích / přeložit
- vyzkoušejte sami na https://www.korpus.cz/slovo-v-kostce/

SyD

- nástroj pro korpusový průzkum variant
- viz https://syd.korpus.cz/
- Sy – varianty v současném jazyce (synchronní část)
  - vyzkoušejte např. *současně* × *najednou* × *naráz*
- D – varianty v průběhu historie (diachronní část)
  - vyzkoušejte *beruška* × *sluněčko*

Morfio

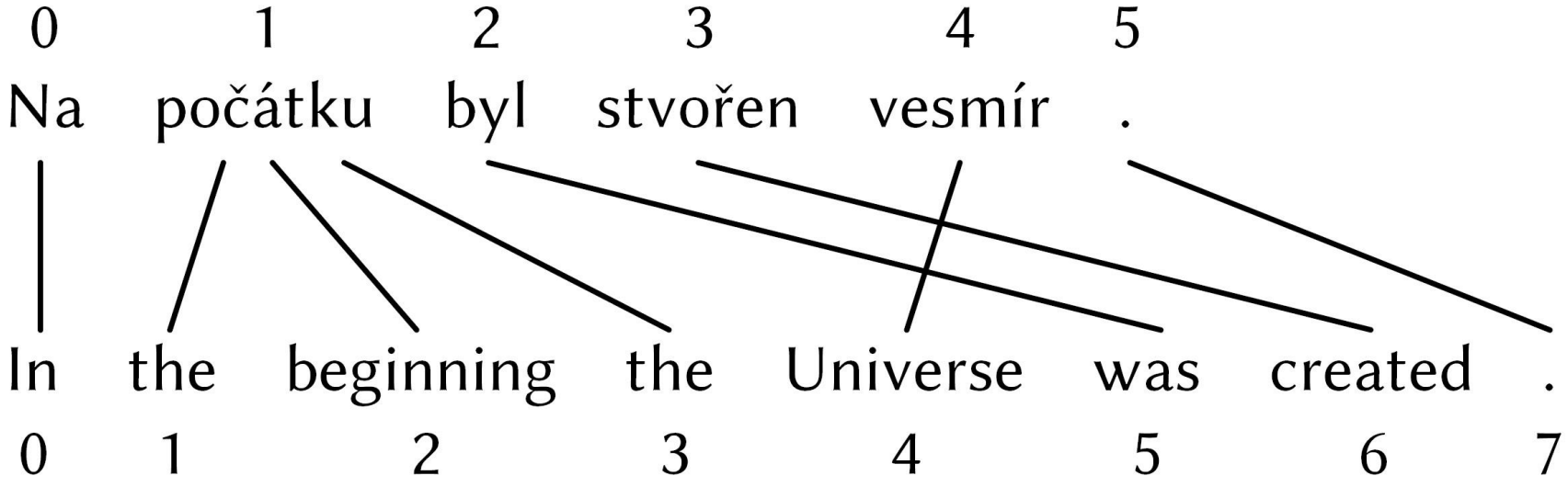
- nástroj pro slovotvornou analýzu
- identifikuje dvojice slov spjaté formálními slovotvornými vztahy
- více informací v hodině věnované slovotvorbě
- https://morfio.korpus.cz/

KWords

- nástroj pro identifikaci klíčových slov
- součást projektu *A Needle in a Haystack*
- zkoumaný text je porovnáván s korpusem (referenční text)
- viz popis aplikace na https://kwords.korpus.cz/

Treq

- databáze překladových ekvivalentů
- vytvořeno automaticky na základě dat z paralelního korpusu *InterCorp* (viz https://intercorp.korpus.cz/)
- snaha zarovnat slovo na slovo:



Převzato z http://wiki.korpus.cz/lib/exe/fetch.php/manualy:carky_gdfa.jpg?cache=

Treq

- zadávání víceslovných výrazů
- možnost využití regulárních výrazů
- nástroj využívaný nejen překladateli
- až 39 jazyků
- vyzkoušejte nástroj na https://treq.korpus.cz/

SKETCH ENGiNE

- komerční korpusový manažer
- firma Lexical Computing, zakladatel Adam Kilgarriff
- 500 korpusů (převážně webových) ve více než 90 jazycích
- https://www.sketchengine.eu/

SKETCH ENGiNE

Přihlášení

- https://auth.sketchengine.eu/#login

Log in

Username or email

Password

LOG IN

Forgot password?

Need help logging in?

or

Institutional login

Sign in with Google

Don't have an account? Sign up

Or try open corpora.

CLICK ON THAT

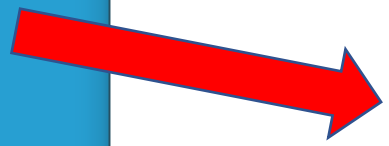
Type an organisation name: Which organisation would you like to sign in with? - Google Chrome

wayf.ukfederation.org.uk/DS-20160527/uk.ds?entityID=https%3A%2F%2Fauth.sketchengine.eu%2Fentity&r...

The Sketch Engine is a Corpus Query System allowing you to research how words behave.

Which organisation would you like to sign in with?

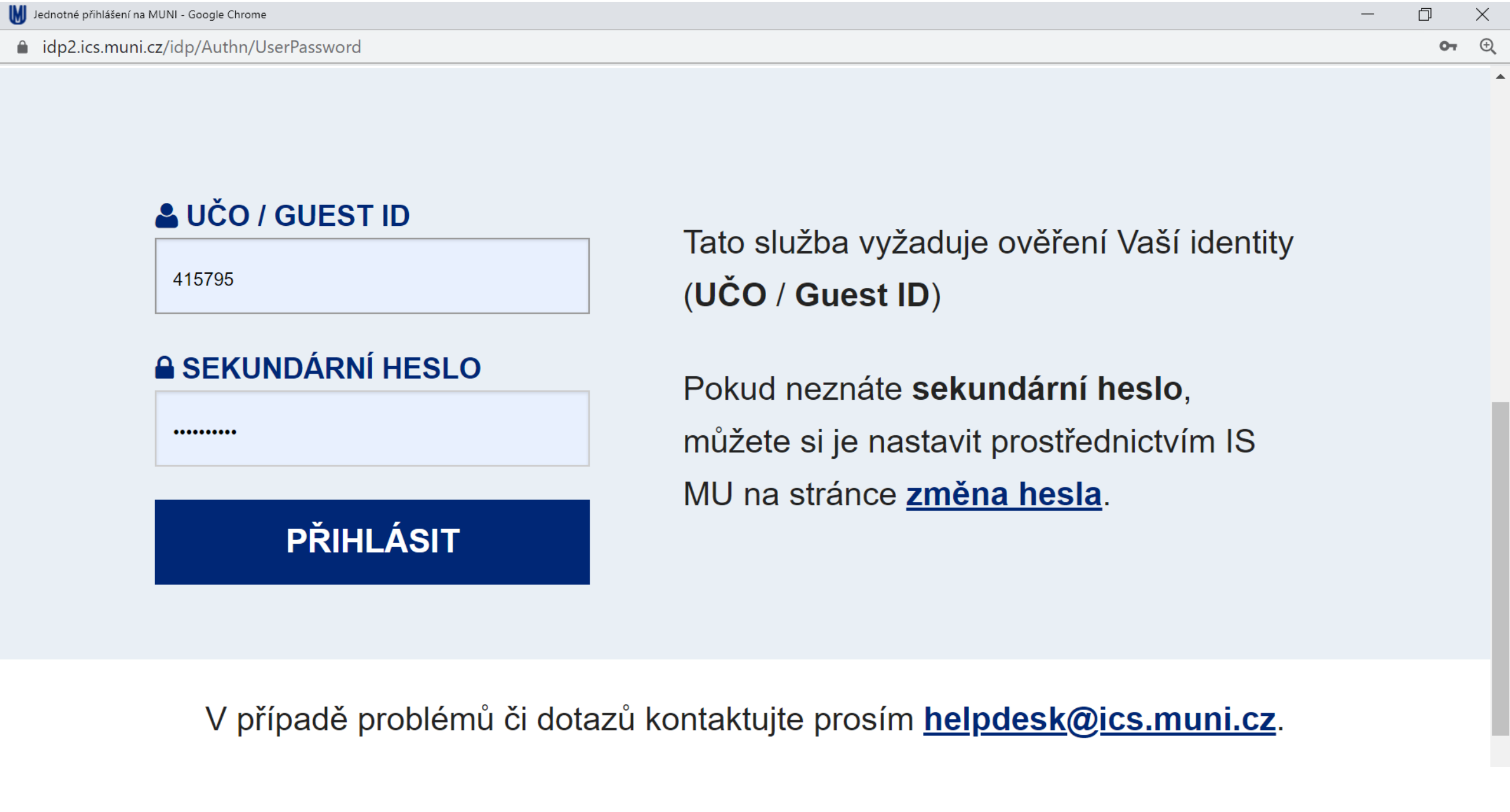
Start typing the name of your organisation (e.g. Anywhere College) in the search box, and options will appear below:



Masarykova univerzita

Continue

or Let me choose from a list


Jednotné přihlášení na MUNI - Google Chrome
idp2.ics.muni.cz/idp/Authn/UserPassword
UČO / GUEST ID
415795
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.

OVLÁDACÍ PANEL

Czech Web 2017 (csTenTen17)

Získat více místa

CZECH WEB 2017 (CSTENTEN17)

INFO O KORPUSU | SPRAVOVAT KORPUS

- **Word Sketch** — Kolokace a kombinace slov
- **Word sketch rozdíly** — Porovnat kolokace dvou slov
- **Tezaurus** — Synonyma a podobná slova
- **Konkordance** — Příklady použití v kontextu
- **Paralelní konkordance** — Hledání překladu
- **Seznam slov** — Frekvenční seznam
- **N-gramy** — Víceslovné výrazy (MWEs)
- **Klíčová slova** — Extrakce terminologie
- **Trendy** — Diachronická analýza, neologismy
- **Slovník jedním kliknutím** — Návrh automatického slovníku

NEDÁVNÉ KORPUSY

NOVÝ KORPUS

Czech Web 2017 (csTenTen17)	Czech	10 502 222 474

SOUČASNÉ VÝSLEDKY | OBLÍBENÉ VÝSLEDKY | ZNAČKOVÁNÍ

hledat psaním

VYBRAT KORPUS

Czech Web 2017 (csTenTen17)

Získat více místa

ZÁKLADNÍ | POKROČILÉ | MÉ KORPUSY | SDÍLENÉ SE MNOU

JAZYKY

Vyberte jazyk a my pro vás zvolíme nejlepší korpus.

ARABIC	DANISH
DUTCH	ENGLISH
FRENCH	GERMAN
CHINESE	ITALIAN
JAPANESE	KOREAN
POLISH	PORTUGUESE
RUSSIAN	SPANISH

Více jazyků

hledat psaním

RYCHLÝ ÚVODNÍ TUTORIÁL

Watch the tutorial

Více info na https://www.sketchengine.eu/guide/

LOG IN | Sign up | FREE Trial | subscribe to news

SKETCH ENGINE

Home | News & Events | Pricing | Guide | About us | Contact

Sketch Engine User Guide

video lessons | face-to-face training | Glossary

general screens

dashboard | select corpus | corpus info page | manage corpus | settings

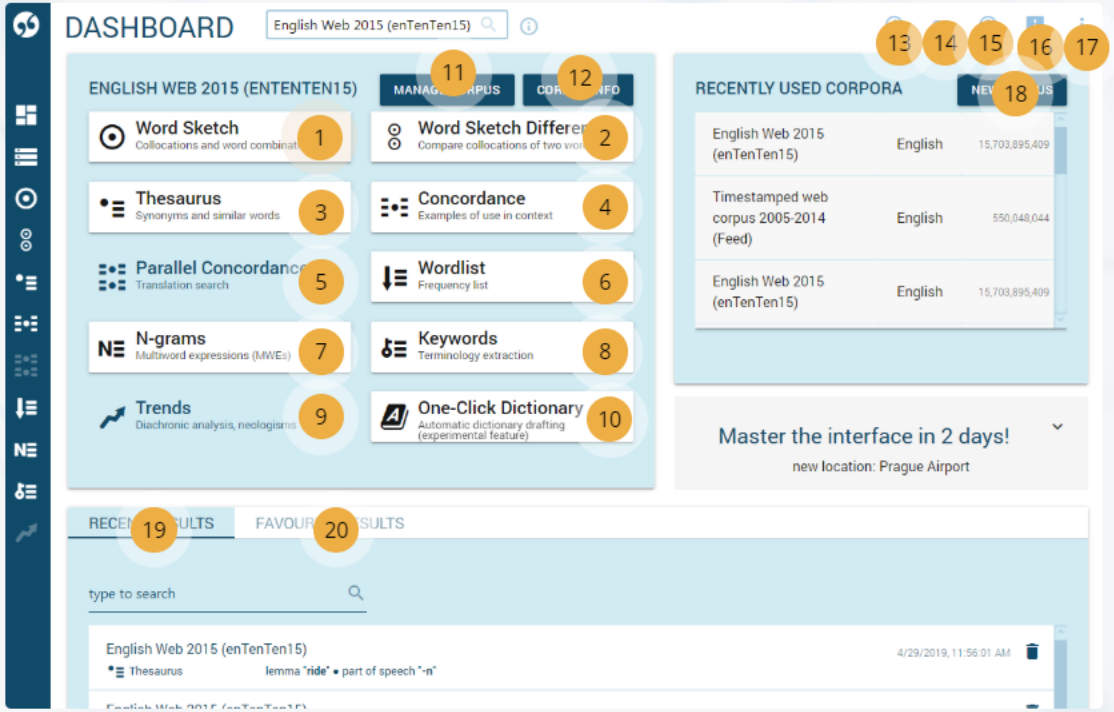
corpus tools

nástroje

word sketch | sketch difference | thesaurus | concordance | parallel concordance | wordlist | n-grams | keywords & terms | bilingual term extraction | trends | OneClick Dictionary

dashboard

Click a place in the dashboard to get help.



Corpus Building & download

build a corpus

Corpora, languages, tagsets

list of languages

Account, subscription, payment

My account and password

Konkordance

- použití daného slova v kontextu
- slovní tvar, lemma, fráze, CQL

KONKORDANCE

Czech Web 2017 (csTenTen17)

Získat více místa

ZÁKLADNÍ | POKROČILÉ | JAK TO FUNGUJE

Typ dotazu

- základní
- **lemma**
- fráze
- slovo
- znak
- CQL

Slovní druh

- jakýkoli
- **podstatné jméno**
- přídavné jméno
- zájmeno
- numeral
- sloveso
- příslovce
- předložka

lemma

koronavirus

A = a ?

Subkorpus

žádný (celý korpus)

Filtrovat kontext

Typy textů ?

OK

KONKORDANCE

Czech Web 2017 (csTenTen17)

Získat více místa

lemma koronavirus 258 (0,02 v mil.)

KWIC

	Detaily	Levý kontext	KWIC	Pravý kontext
1	wikipedia.org	í původci nemoci napadají jen jisté zvířecí druhy. </s><s>	**Koronavirus** k1gInSc1/Koronavirus	přenosné gastroenteritidy (TGEV) se objevuje u prasat a z
2	wikipedia.org	nes, holubů, a japonských křepelek. </doc></s><s><doc>	**Koronavirus** k1gInSc1/Koronavirus	(Coronavirus) je společné označení pro čtyři rody virů obs
3	wikipedia.org	esnila, že původní údaje o počtu osob zemřelých na tento	**koronavirus** k1gInSc4/koronavirus	byly podhodnocené, a že ve skutečnosti zemřelo 282 lidí (
4	wikipedia.org	nc Felipe.[6] </s><s> Saúdskoarabské úřady oznámily, že	**koronavirus** k1gInSc1/koronavirus	MERS si v zemi vyžádal již 282 obětí. [7] </s><s> Za bojk
5	tribune.cz	oho u 22 pikornavirus, u sedmi metapneumovirus, u osmi	**koronavirus** k1gInSc1/koronavirus	, u dvou influenza virus typu A/B a parainfluenza virus a u
6	zet.cz	yžádala 46 životů, nakaženo bylo celkem 94 lidí. </s><s>	**Koronavirus** k1gInSc1/Koronavirus	MERS může vyvolat kašel, vážné dýchací potíže, horečku
7	zet.cz	otilátky, o kterých si myslíme, že jsou specifické právě pro	**koronavirus** k1gInSc4/koronavirus	MERS nebo jemu podobný virus, v krvi jednohrbých velblc
8	tribune.cz	ně </s><s> V Česku se objevil druhý případ podezření na	**koronavirus** k1gInSc4/koronavirus	MERS. </s><s> Muž z Ostravy byl hospitalizován na infek
9	tribune.cz	> V České republice se objevil druhý případ podezření na	**koronavirus** k1gInSc4/koronavirus	MERS. </s><s> Muž z Ostravy byl hospitalizován na infek
10	tribune.cz	zdravotnický systém ji zvládne. </s><s> Infekci způsobuje	**koronavirus** k1gInSc1/koronavirus	z téhož kmene, jako je virus zodpovědný za onemocnění S
11	euro.cz	íkům otvírá už v deset hodin. </doc></s><s><doc> WHO:	**Koronavirus** k1gMnSc1/Koronavirus	MERS zatím není celosvětovou hrozbou </s><s> Přestože
12	euro.cz	dýchacích obtíží i selhání ledvin. </s><s> Celosvětově na	**koronavirus** k1gInSc4/koronavirus	MERS zemřelo už 152 lidí, naprostá většina v Saúdské Ar
13	aktualne.cz	sobuje symptomy podobné nákaze virem SARS. </s><s>	**Koronavirus** k1gMnSc1/Koronavirus	je označení pro čtyři rody virů, které způsobují méně záva
14	brnenskadrbna.cz	áb </s><s> V Česku se objevil druhý případ podezření na	**koronavirus** k1gInSc4/koronavirus	MERS. </s><s> Třiatřicetiletý pracovník brněnské cestovn
15	euro.cz	zahlédnete, kontaktujte policii. </doc></s><s><doc> Nový	**koronavirus**	je dle WHO přenosný kontaktem s nakaženou osobou </s

KONKORDANCE

Czech Web 2017 (csTenTen17)

Získat více místa

lemma koronavirus 258 (0,02 v mil.)

KWIC

	Detaily	Levý kontext	KWIC	Pravý kontext
1	wikipedia.org	í původci nemoci napadají jen jisté zvířecí druhy. </s><s>	**Koronavirus** k1glnSc1/Koronavirus	přenosné gastroenteritidy (TGEV) se objevuje u prasat a z
2	wikipedia.org	ıes, holubů, a japonských křepelek. </doc></s><s><doc>	**Koronavirus** k1glnSc1/Koronavirus	(Coronavirus) je společné označení pro čtyři rody virů obs
3	wikipedia.org	esnila, že původní údaje o počtu osob zemřelých na tento	**koronavirus** k1glnSc4/koronavirus	byly podhodnocené, a že ve skutečnosti zemřelo 282 lidí (
4	wikipedia.org	ıc Felipe.[6] </s><s> Saúdskoarabské úřady oznámily, že	**koronavirus** k1glnSc1/koronavirus	MERS si v zemi vyžádal již 282 obětí. [7] </s><s> Za bojk
5	tribune.cz	oho u 22 pikornavirus, u sedmi metapneumovirus, u osmi	**koronavirus** k1glnSc1/koronavirus	, u dvou influenza virus typu A/B a parainfluenza virus a u
6	zet.cz			otíže, horečku
7	zet.cz			ohrbých velblc
8	tribun			zován na infek
9	tribun			zován na infek
10	tribun			onemocnění s
11	euro.c			><s> Přestože
12	euro.c			v Saúdské Ar
13	aktua			ují méně záva
14	brnen			ěnské cestovn
15	euro.c			ou osobou </s

virus SARS, se začala šířit z Blízkého východu a od loňska si vyžádala 46 životů, nakaženo bylo celkem 94 lidí. </s> <s> Choroba, která připomíná virus SARS, se začala šířit z Blízkého východu a od loňska si vyžádala 46 životů, nakaženo bylo celkem 94 lidí. </s> <s> **Koronavirus** MERS může vyvolat kašel, vážné dýchací potíže, horečku, zápal plic a údajně i selhání ledvin. </s><s> Úmrtnost nakažených pacientů přesahuje 50 procent, což je mnohem víc než v případě příbuzného koronaviru SARS. </s> <s> Důkazy pro přenos nemoci mezi lidmi existovaly, odborníci se ale vždy domnívali

Word Sketch

- tzv. slovní profily
- sdružují kolokace slov na základě gramatických relací (podmět, přísudek, atribut, …)
- hodnoty frekvence a skóre

WORD SKETCH

Czech Web 2017 (csTenTen17)

Získat více místa

ZÁKLADNÍ | POKROČILÉ | JAKO SEZNAM | JAK TO FUNGUJE

Hledání ?

krásný

Subkorpus ?

žádný (celý korpus)

Slovní druh ?

- **auto**
- adjective
- adverb
- noun
- pronoun
- verb
- numeral
- preposition

Minimální frekvence ?

auto

Minimální skóre ?

0

Přeložit

Typy textů ?

OK

WORD SKETCH

Czech Web 2017 (csTenTen17)

Získat více místa

krásný as přídavné jméno 3 354 950×

modifiers of "krásný"

moc moc krásné	39 843	9,1
opravdu opravdu krásné	26 054	8,27
nejen nejen krásné	7 957	7,89
uvěřitelně neuvěřitelně krásné	2 841	7,59
tak tak krásné	36 884	7,34
opět opět krásné	4 082	7,18
přirozeně přirozeně krásná	1 555	7,07
skutečně skutečně krásné	5 535	7,05
strašně strašně krásný	2 501	7,03
plno plno krásných	1 331	6,97
tu tu krásné	5 362	6,93
spoustu spoustu krásných	1 654	6,93

nouns modified by "krásný"

výhled s krásným výhledem	64 223	9,53
příroda krásné přírodě	42 731	8,96
počasí krásné počasí	43 816	8,85
den krásný den	108 899	8,77
pláž krásné pláže	33 314	8,56
žena krásné ženy	33 172	8,27
prostředí v krásném prostředí	36 253	7,94
fotka krásné fotky	20 955	7,89
zážitek krásný zážitek	24 066	7,86
místo krásné místo	38 407	7,63
dívka krásné dívky	17 267	7,63
dárek krásný dárek	19 731	7,6

"krásný" and/or ...

zdravý krásné a zdravé	8 471	9,46
klidný krásné a klidné	6 465	9,22
pohodový krásné a pohodové	4 040	8,89
zajímavý krásné a zajímavé	5 516	8,64
mladý mladá a krásná	4 399	8,51
čistý krásné a čisté	3 742	8,3
milý krásné a milé	2 486	7,89
originální krásné a originální	2 238	7,82
bohatý krásné a bohaté	2 136	7,68
elegantní krásné a elegantní	1 924	7,65
kvalitní krásné a kvalitní	2 999	7,59
veliký největší a nejkrásnější	2 862	7,52

... is "krásný"

štěstí štěstí je krásná věc	2 323	9,53
život Život je krásný	3 072	9,46
fotka fotky jsou krásné	1 048	8,2
počasí Počasí bylo krásné	1 020	8,16
žena žena je krásná	1 038	8,08
svět svět je krásný	1 107	7,92
láska Láska je krásná	723	7,74
okolí V okolí je krásná	727	7,73
den den je krásný	1 014	7,62
pláž Pláž je krásná	572	7,39
hotel Hotel je krásný	619	7,27
Praha Praha je krásná	537	7,02

words before "krásný"

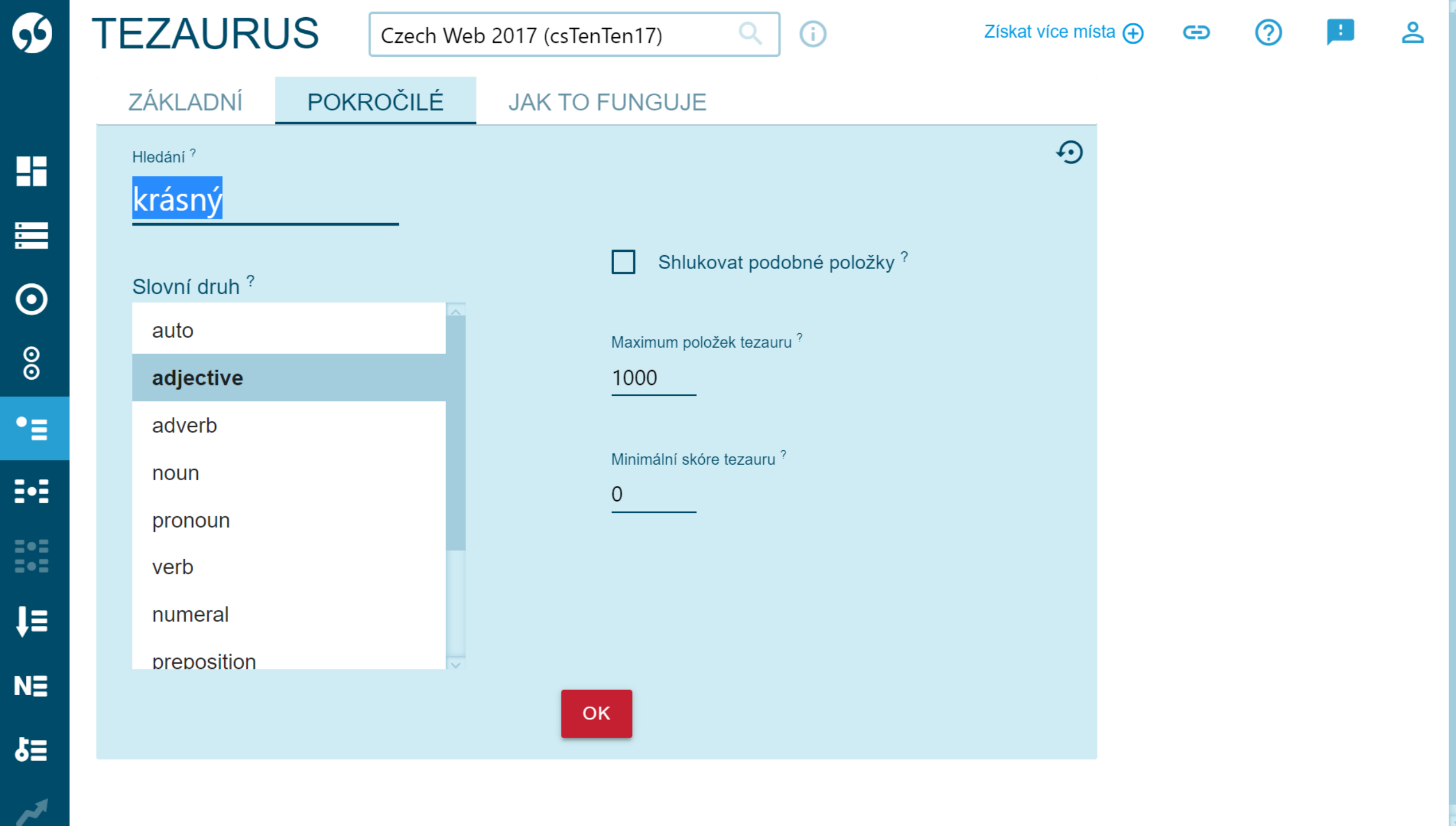
fakt je fakt krásná	2 034	7,96
vskutku vskutku krásné	714	6,94
bezesporu bezesporu nejkrásnější	587	6,79
prostě prostě krásná	1 901	5,59
opravdu opravdu krásný	205	5,51
přece přece krásné	1 149	5,46
mimochodem mimochodem krásné	236	5,41
prý prý nejkrásnější	1 029	5,34
zase zase krásný	1 930	5,15
také také krásné	9 062	5,1
jistě jistě krásné	1 022	5,09
ra book of ra nejkrásnější	440	5,08

words after "krásný"

ba krásný ba	34	5,01
jednou krásné jednou	48	4,79
ch z nejkrásnější ch	18	4,17
akorát krásné akorát	14	4,17
ach všechno zdá se krásnější Ach	14	3,61
pr krásné PR	45	3,59
jaký nejkrásnější jaký jsem	19	3,57
právě krásná právě	440	3,51
co to nejkrásnější co	843	3,49
kde Krásný Kde	15	2,93
ahoj krásné ahoj	23	2,85
bezesporu nejkrásnější bezesporu patří	13	2,72

Tezaurus

- The thesaurus in Sketch Engine is an automatically generated list of synonyms or words belonging to the same category (semantic field). The list is produced based on the context in which the words appear in the selected corpus. Only nouns, adjectives, verbs and adverbs are supported in most corpora


TEZAURUS
Czech Web 2017 (csTenTen17)
Získat více místa
ZÁKLADNÍ
POKROČILÉ
JAK TO FUNGUJE
Hledání ?
krásný
Slovní druh ?
auto
adjective
adverb
noun
pronoun
verb
numeral
preposition
Shlukovat podobné položky ?
Maximum položek tezauru ?
1000
Minimální skóre tezauru ?
0
OK

krásný as přídavné jméno 3 354 950×

	Word	Frekvence ?	
1	nádherný	1 063 308	•••
2	hezký	1 286 906	•••
3	pěkný	1 677 821	•••
4	úžasný	1 362 352	•••
5	zajímavý	4 050 171	•••
6	skvělý	3 022 707	•••
7	příjemný	2 802 641	•••
8	plný	3 222 664	•••
9	dobrý	20 585 983	•••
10	překrásný	158 798	•••

n-gramy

- sled po sobě jdoucích položek z dané posloupnosti
- unigram, bigram, trigram…
- kolokace vs. n-gram

The office building was demolished yesterday.

5 bigramů: *the office, office building, building was, was demolished, demolished yesterday*

2 kolokace: *office building, to demolish a building*

N-GRAMY

Czech Web 2017 (csTenTen17)

Získat více místa

ZÁKLADNÍ | POKROČILÉ | JAK TO FUNGUJE

Délka n-gramu ? 2 3 4 5 6

Zanořené n-gramy ? Zahrnout neslovní výrazy

A = a ? mimo tato slova: ?

Atribut ?

slovo

Frekvence min. ? 5

Frekvence max. ? 0

Subkorpus ?

žádný (celý korpus)

Klíčové n-gramy

Additional criteria

- vše
- začínající písmeny
- končící písmeny
- **obsahující písmena**
- začínající slovem
- obsahující slovo
- končící slovem
- odpovídá regulárnímu výrazu

černý

PŘIDAT TUTO PODMÍNKU

Typy textů ?

OK

N-GRAMY

Czech Web 2017 (csTenTen17)

Získat více místa

2–3-grams

slovo

(položky: **1 630**, celková frekvence: **33 088**)

	Word	↓ Počet ?
1	a černý	1 198
2	černý čaj	904
3	je černý	819
4	černý a	644
5	černý humor	535
6	na černý	447
7	černý trh	422
8	černý rybíz	401
9	černý pepř	309
10	jako černý	297
11	černý toner	280
12	černý kašel	275
13	má černý	255
14	nebo černý	232
15	černý s	219
16	černý nebo	218
17	černý kůň	209
18	černý je	191
19	černý bez	183
20	ten černý	175
21	i černý	171
22	černý pruh	164
23	se černý	161
24	velký černý	155
25	čáp černý	154
26	černý se	154
27	byl černý	149
28	že černý	141
29	černý dým	137
30	černý pátek	135
31	celý černý	134
32	pro černý	128
33	černý kouř	125
34	černý na	122
35	černý pasažér	119
36	černý plech	116
37	černý povrch	110
38	černý pes	106
39	černý lak	106
40	jen černý	105
41	bez černý	103
42	až černý	103
43	to černý	98
44	černý plast	97
45	datel černý	97
46	černý kontinent	93
47	černý trh s	92
48	černý jako	90
49	o černý	88
50	černý den	87

Máte přístup pouze k 1 000 položkám. Získat více

Řádků na stránku 50 1–50 z 1 000 1 / 20

SkELL

- webová aplikace pro učení jazyka na základě dat z korpusu
- česká verze: https://csskell.sketchengine.co.uk/run.cgi/skell
- verze pro angličtinu, ruštinu, italštinu, němčinu nebo estonštinu

Děkuji za pozornost