CZECH NATIONAL
CORPUS

CZECH NATIONAL
CORPUS

Introduction to Text Corpora and Their Applications

Corpora in lexical studies and lexicography

Lucie Chlumská, Ph.D.

lucie.chlumska@korpus.cz

OUTLINE:

1. LECTURE

- revision: lexicography b.c. and from 1990s onwards
- corpus-based lexical studies focusing mainly on:
  - frequency
  - collocations

2. SEMINAR

- reading (Hans Lindquist): *Looking for lexis*
- collocations in a dictionary: in search of meaning and collocability

LECTURE

Lexicography b.c.

The beginnings

First attempts to collect data similar to corpora (before 1960s) were made in the following areas:

- biblical and literary studies
- lexicography
- dialect studies
- language education studies
- grammatical studies

Pre-corpus lexicography

- as early as 17th century
- Samuel Johnson recorded on slips of paper a large corpus of sentences from ‘writers of the first reputation’ to illustrate meanings and uses of English words in his *Dictionary of the English Language*
  - Johnson worked with 6 assistants to assemble over 150,000 illustrative citations for the app. 40,000 headword entries
- similarly, *Oxford English Dictionary* (*OED*) also corpus-based
  - twelfth and final volume published in 1928
  - 71 years of sustained work on a corpus of the canon of mainly literary written English from about AD 1000
  - 2,000 volunteer readers collected about five million citations amounting to 50 million words to illustrate 414,825 entries

Pre-corpus lexicography

- parallel to the work on the second edition of *OED* in the latter part of 19th century, another great corpus of citations was being assembled to support the third edition of Noah Webster's *An American Dictionary of the English Language*

  - in 1961, the third edition of Webster's *New International Dictionary* had available a corpus of over 10 million citation slips
  - probably the last major English dictionary to be completed without and electronic database...

Corpus-based lexicography

Benefits of using corpora

- advantages: large amount of data, annotation & mark-up

- five changes brought about by corpora to dictionaries:

1. an emphasis on frequency;
2. an emphasis on collocation and phraseology;
3. an emphasis on variation;
4. an emphasis on lexis in grammar;
5. an emphasis on authenticity.

Corpus-based dictionaries

- COBUILD = Collins Birmingham University International Language Database
  - since the 1980, led by John Sinclair
  - Collins Corpus > Bank of English

  - Collins Cobuild English Language Dictionary
    - 1st edition 1987, 2nd edition 1995
    - defines over 70,000 words, giving priority to the most frequent
    - definitions are generally supported by examples of usage taken from the Cobuild corpus

Corpus-based dictionaries

- Longman Dictionary of Contemporary English
  - first published in 1978
  - project guided by Randolph Quirk
  - intended primarily for the foreign user
  - definitions are always written using simpler terms than the words they describe (core vocabulary of 2000 most frequent words used in definitions)

  - 3rd edition 1995
  - more user-friendly
  - 2 300 words illustrated, 24 pages in full colour

Collocations

Collocation

- collocation = a co-occurrence relationship between two words: a node word and its collocate
  - based on statistics (frequency and probability)
  - association measures
    - t-score, MI-score, LogDice etc.
    - no measure is perfect...
- colligation = a collocation of a node word with a particular grammatical class of words

> *What collocation is on a lexical level of analysis, colligation is on a syntactic level. The term does not refer to the repeated combination of concrete word forms but to the way in which word classes co-occur or keep habitual company in an utterance*
>
> Ute Römer

Collocation

- J. R. Firth (1957): term *collocation* (Latin collocare = place together)

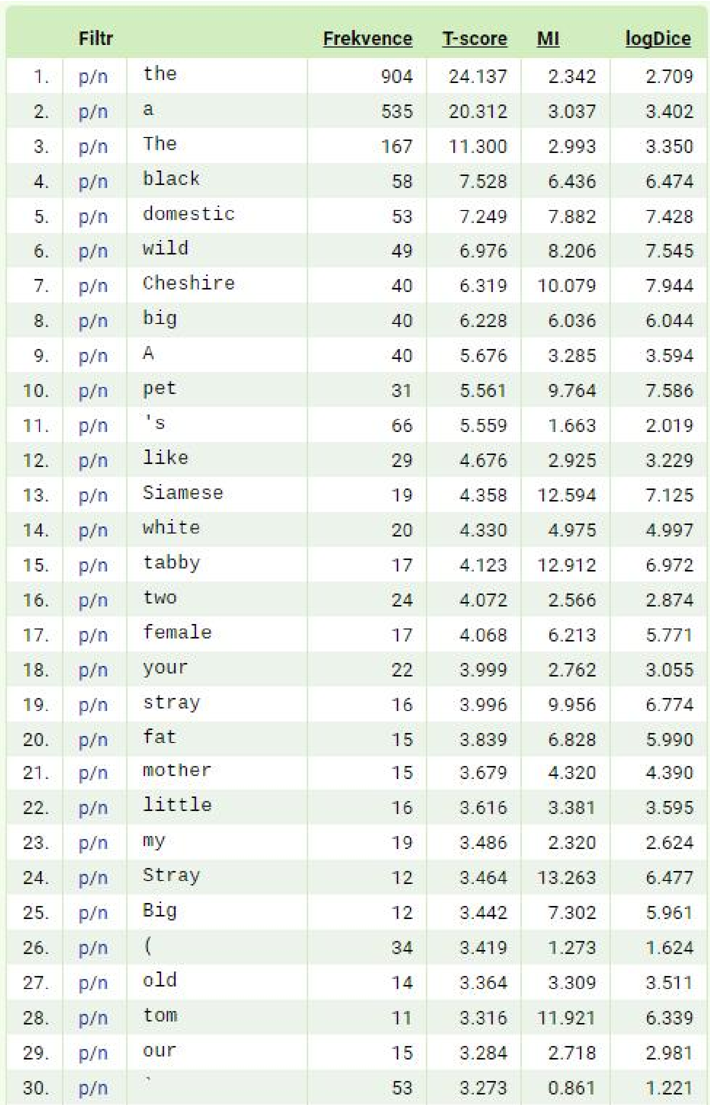
„Collocations of a given word are statements of the habitual or customary places of that word.“ (1968: 181)

- Greenbaum (1974): intuition as a poor guide to collocation
  - introspection-based elicitation experiments > people disagree on collocations, because *„each of us has only a partial knowledge of the language, we have prejudices and preferences, our memory is week, we tend to notice unusual words and structures but often overlook the ordinary ones“*
    (Krishnamurthy 200: 32-33)

- Partington (1998): *„there is no total agreement among native speakers as to which collocations are acceptable and which are not“*

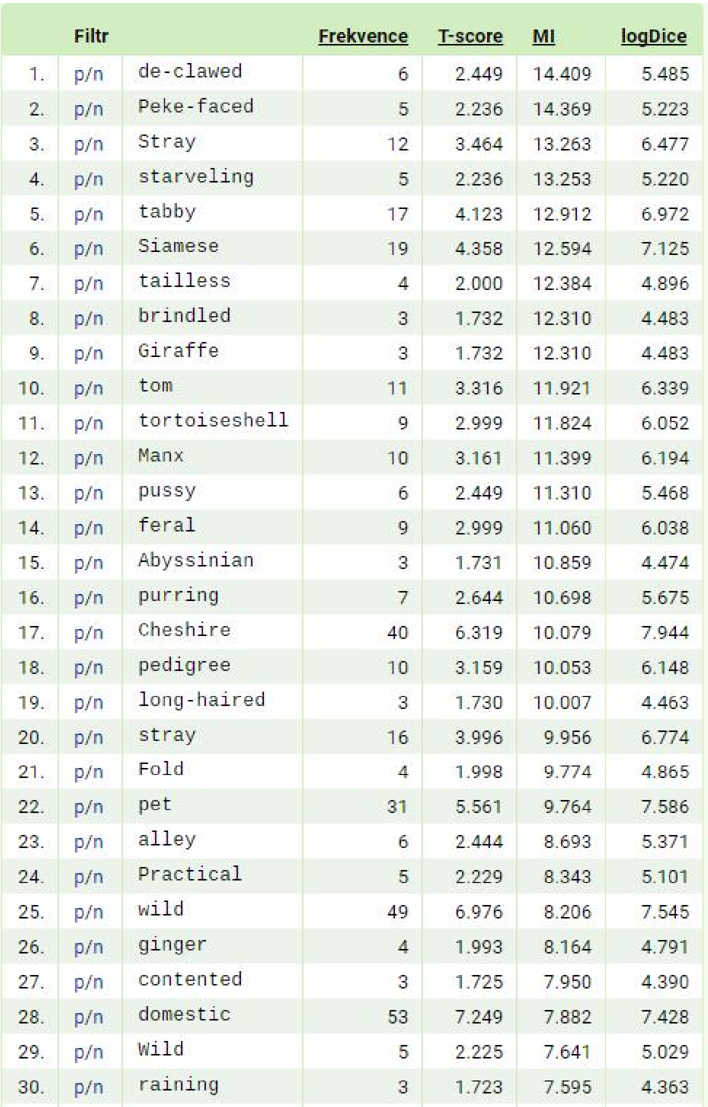
Association measures

t-score



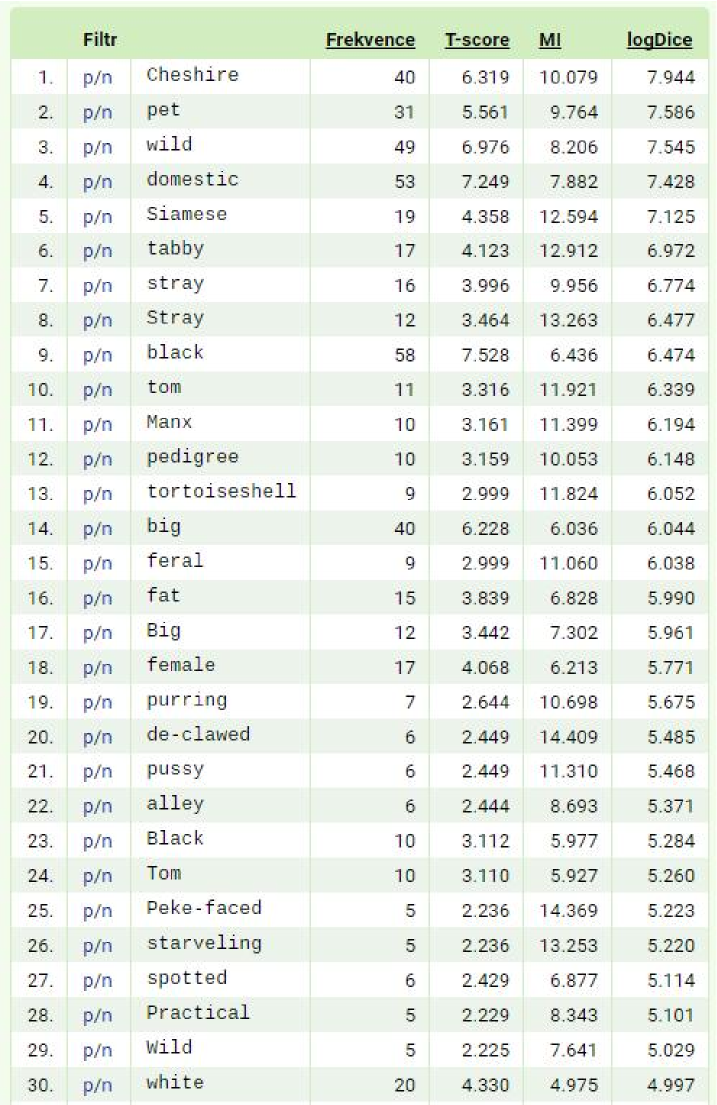
	Filtr		Frekvence	T-score	MI	logDice
1.	p/n	the	904	24.137	2.342	2.709
2.	p/n	a	535	20.312	3.037	3.402
3.	p/n	The	167	11.300	2.993	3.350
4.	p/n	black	58	7.528	6.436	6.474
5.	p/n	domestic	53	7.249	7.882	7.428
6.	p/n	wild	49	6.976	8.206	7.545
7.	p/n	Cheshire	40	6.319	10.079	7.944
8.	p/n	big	40	6.228	6.036	6.044
9.	p/n	A	40	5.676	3.285	3.594
10.	p/n	pet	31	5.561	9.764	7.586
11.	p/n	's	66	5.559	1.663	2.019
12.	p/n	like	29	4.676	2.925	3.229
13.	p/n	Siamese	19	4.358	12.594	7.125
14.	p/n	white	20	4.330	4.975	4.997
15.	p/n	tabby	17	4.123	12.912	6.972
16.	p/n	two	24	4.072	2.566	2.874
17.	p/n	female	17	4.068	6.213	5.771
18.	p/n	your	22	3.999	2.762	3.055
19.	p/n	stray	16	3.996	9.956	6.774
20.	p/n	fat	15	3.839	6.828	5.990
21.	p/n	mother	15	3.679	4.320	4.390
22.	p/n	little	16	3.616	3.381	3.595
23.	p/n	my	19	3.486	2.320	2.624
24.	p/n	Stray	12	3.464	13.263	6.477
25.	p/n	Big	12	3.442	7.302	5.961
26.	p/n	(	34	3.419	1.273	1.624
27.	p/n	old	14	3.364	3.309	3.511
28.	p/n	tom	11	3.316	11.921	6.339
29.	p/n	our	15	3.284	2.718	2.981
30.	p/n	`	53	3.273	0.861	1.221

MI-score (mutual information)



	Filtr		Frekvence	T-score	MI	logDice
1.	p/n	de-clawed	6	2.449	14.409	5.485
2.	p/n	Peke-faced	5	2.236	14.369	5.223
3.	p/n	Stray	12	3.464	13.263	6.477
4.	p/n	starveling	5	2.236	13.253	5.220
5.	p/n	tabby	17	4.123	12.912	6.972
6.	p/n	Siamese	19	4.358	12.594	7.125
7.	p/n	tailless	4	2.000	12.384	4.896
8.	p/n	brindled	3	1.732	12.310	4.483
9.	p/n	Giraffe	3	1.732	12.310	4.483
10.	p/n	tom	11	3.316	11.921	6.339
11.	p/n	tortoiseshell	9	2.999	11.824	6.052
12.	p/n	Manx	10	3.161	11.399	6.194
13.	p/n	pussy	6	2.449	11.310	5.468
14.	p/n	feral	9	2.999	11.060	6.038
15.	p/n	Abyssinian	3	1.731	10.859	4.474
16.	p/n	purring	7	2.644	10.698	5.675
17.	p/n	Cheshire	40	6.319	10.079	7.944
18.	p/n	pedigree	10	3.159	10.053	6.148
19.	p/n	long-haired	3	1.730	10.007	4.463
20.	p/n	stray	16	3.996	9.956	6.774
21.	p/n	Fold	4	1.998	9.774	4.865
22.	p/n	pet	31	5.561	9.764	7.586
23.	p/n	alley	6	2.444	8.693	5.371
24.	p/n	Practical	5	2.229	8.343	5.101
25.	p/n	wild	49	6.976	8.206	7.545
26.	p/n	ginger	4	1.993	8.164	4.791
27.	p/n	contented	3	1.725	7.950	4.390
28.	p/n	domestic	53	7.249	7.882	7.428
29.	p/n	Wild	5	2.225	7.641	5.029
30.	p/n	raining	3	1.723	7.595	4.363

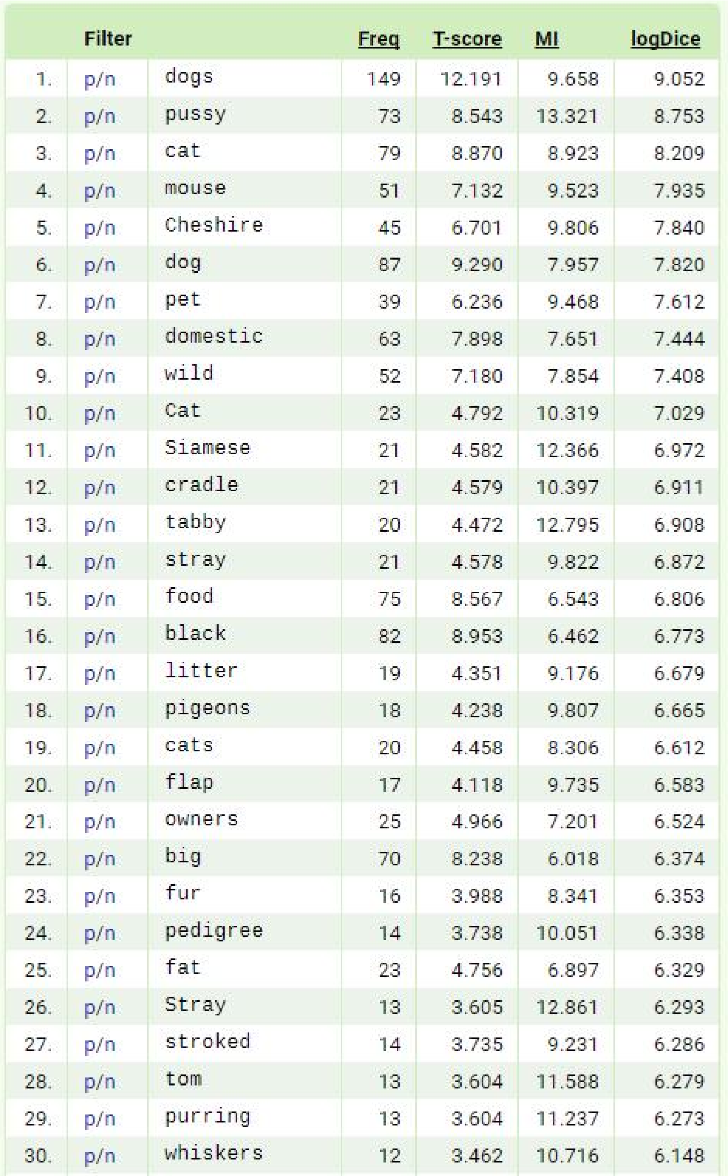
logDice



	Filtr		Frekvence	T-score	MI	logDice
1.	p/n	Cheshire	40	6.319	10.079	7.944
2.	p/n	pet	31	5.561	9.764	7.586
3.	p/n	wild	49	6.976	8.206	7.545
4.	p/n	domestic	53	7.249	7.882	7.428
5.	p/n	Siamese	19	4.358	12.594	7.125
6.	p/n	tabby	17	4.123	12.912	6.972
7.	p/n	stray	16	3.996	9.956	6.774
8.	p/n	Stray	12	3.464	13.263	6.477
9.	p/n	black	58	7.528	6.436	6.474
10.	p/n	tom	11	3.316	11.921	6.339
11.	p/n	Manx	10	3.161	11.399	6.194
12.	p/n	pedigree	10	3.159	10.053	6.148
13.	p/n	tortoiseshell	9	2.999	11.824	6.052
14.	p/n	big	40	6.228	6.036	6.044
15.	p/n	feral	9	2.999	11.060	6.038
16.	p/n	fat	15	3.839	6.828	5.990
17.	p/n	Big	12	3.442	7.302	5.961
18.	p/n	female	17	4.068	6.213	5.771
19.	p/n	purring	7	2.644	10.698	5.675
20.	p/n	de-clawed	6	2.449	14.409	5.485
21.	p/n	pussy	6	2.449	11.310	5.468
22.	p/n	alley	6	2.444	8.693	5.371
23.	p/n	Black	10	3.112	5.977	5.284
24.	p/n	Tom	10	3.110	5.927	5.260
25.	p/n	Peke-faced	5	2.236	14.369	5.223
26.	p/n	starveling	5	2.236	13.253	5.220
27.	p/n	spotted	6	2.429	6.877	5.114
28.	p/n	Practical	5	2.229	8.343	5.101
29.	p/n	Wild	5	2.225	7.641	5.029
30.	p/n	white	20	4.330	4.975	4.997

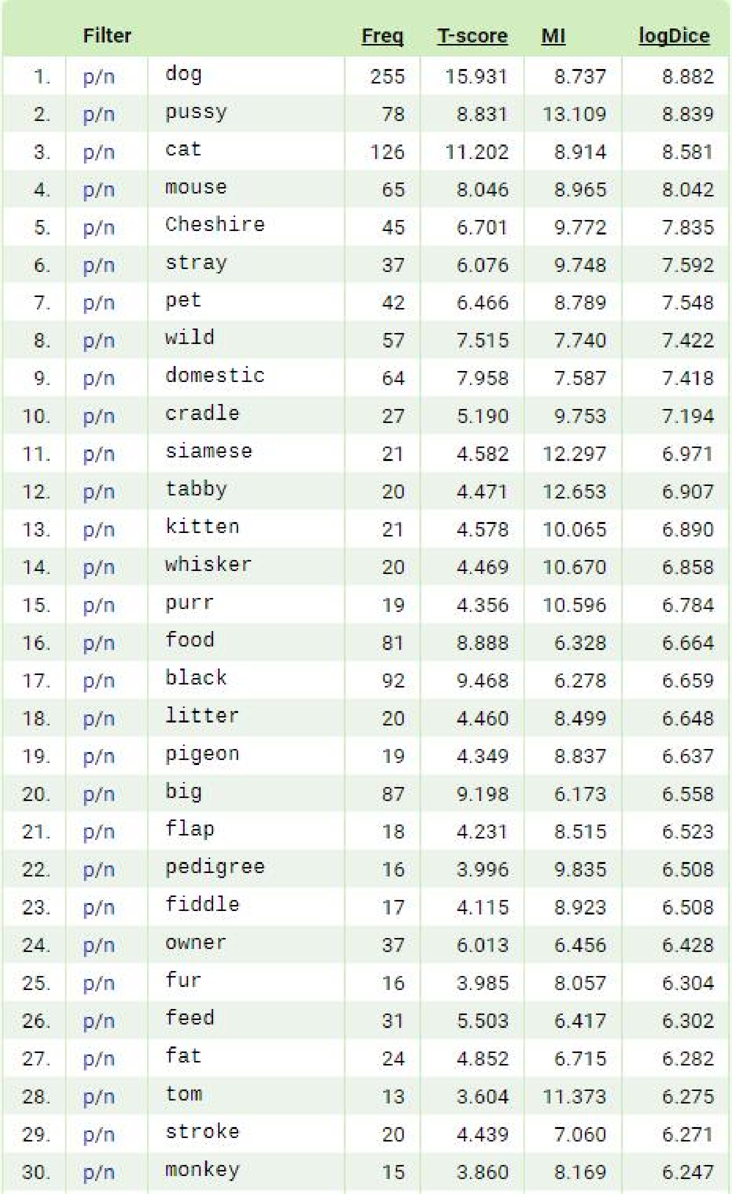
word or lemma collocate?

word (logDice, -3 +3)



	Filter		Freq	T-score	MI	logDice
1.	p/n	dogs	149	12.191	9.658	9.052
2.	p/n	pussy	73	8.543	13.321	8.753
3.	p/n	cat	79	8.870	8.923	8.209
4.	p/n	mouse	51	7.132	9.523	7.935
5.	p/n	Cheshire	45	6.701	9.806	7.840
6.	p/n	dog	87	9.290	7.957	7.820
7.	p/n	pet	39	6.236	9.468	7.612
8.	p/n	domestic	63	7.898	7.651	7.444
9.	p/n	wild	52	7.180	7.854	7.408
10.	p/n	Cat	23	4.792	10.319	7.029
11.	p/n	Siamese	21	4.582	12.366	6.972
12.	p/n	cradle	21	4.579	10.397	6.911
13.	p/n	tabby	20	4.472	12.795	6.908
14.	p/n	stray	21	4.578	9.822	6.872
15.	p/n	food	75	8.567	6.543	6.806
16.	p/n	black	82	8.953	6.462	6.773
17.	p/n	litter	19	4.351	9.176	6.679
18.	p/n	pigeons	18	4.238	9.807	6.665
19.	p/n	cats	20	4.458	8.306	6.612
20.	p/n	flap	17	4.118	9.735	6.583
21.	p/n	owners	25	4.966	7.201	6.524
22.	p/n	big	70	8.238	6.018	6.374
23.	p/n	fur	16	3.988	8.341	6.353
24.	p/n	pedigree	14	3.738	10.051	6.338
25.	p/n	fat	23	4.756	6.897	6.329
26.	p/n	Stray	13	3.605	12.861	6.293
27.	p/n	stroked	14	3.735	9.231	6.286
28.	p/n	tom	13	3.604	11.588	6.279
29.	p/n	purring	13	3.604	11.237	6.273
30.	p/n	whiskers	12	3.462	10.716	6.148

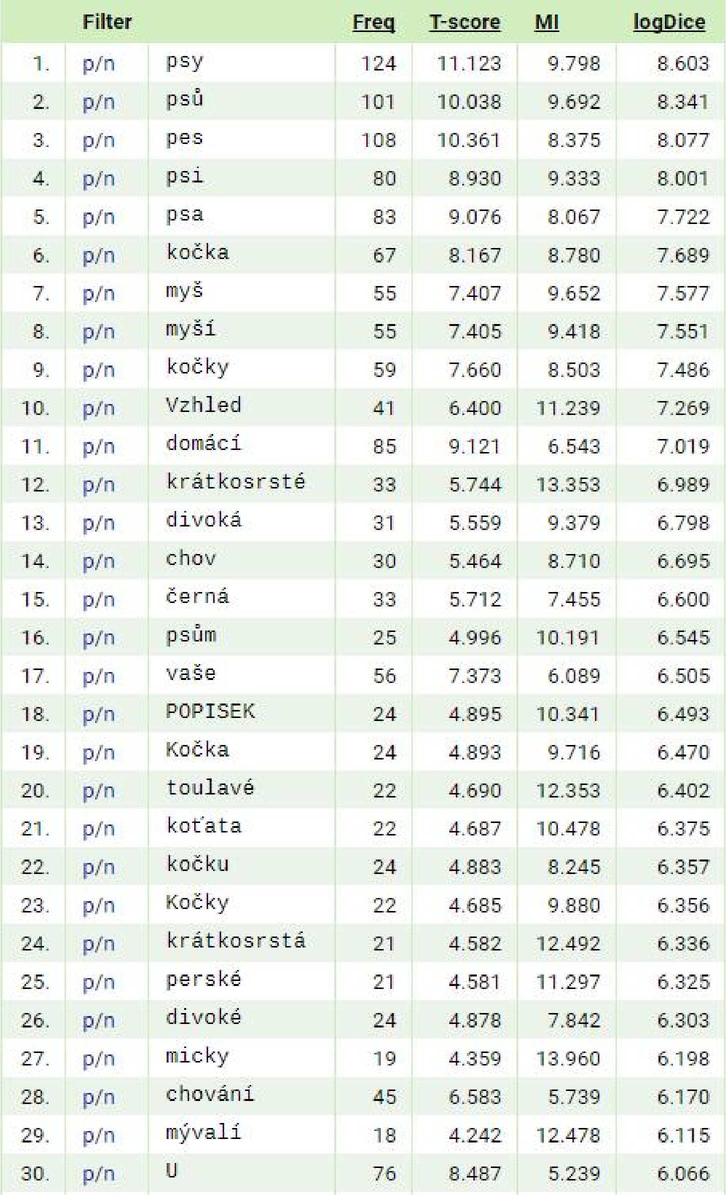
lemma (logDice, -3 +3)



	Filter		Freq	T-score	MI	logDice
1.	p/n	dog	255	15.931	8.737	8.882
2.	p/n	pussy	78	8.831	13.109	8.839
3.	p/n	cat	126	11.202	8.914	8.581
4.	p/n	mouse	65	8.046	8.965	8.042
5.	p/n	Cheshire	45	6.701	9.772	7.835
6.	p/n	stray	37	6.076	9.748	7.592
7.	p/n	pet	42	6.466	8.789	7.548
8.	p/n	wild	57	7.515	7.740	7.422
9.	p/n	domestic	64	7.958	7.587	7.418
10.	p/n	cradle	27	5.190	9.753	7.194
11.	p/n	siamese	21	4.582	12.297	6.971
12.	p/n	tabby	20	4.471	12.653	6.907
13.	p/n	kitten	21	4.578	10.065	6.890
14.	p/n	whisker	20	4.469	10.670	6.858
15.	p/n	purr	19	4.356	10.596	6.784
16.	p/n	food	81	8.888	6.328	6.664
17.	p/n	black	92	9.468	6.278	6.659
18.	p/n	litter	20	4.460	8.499	6.648
19.	p/n	pigeon	19	4.349	8.837	6.637
20.	p/n	big	87	9.198	6.173	6.558
21.	p/n	flap	18	4.231	8.515	6.523
22.	p/n	pedigree	16	3.996	9.835	6.508
23.	p/n	fiddle	17	4.115	8.923	6.508
24.	p/n	owner	37	6.013	6.456	6.428
25.	p/n	fur	16	3.985	8.057	6.304
26.	p/n	feed	31	5.503	6.417	6.302
27.	p/n	fat	24	4.852	6.715	6.282
28.	p/n	tom	13	3.604	11.373	6.275
29.	p/n	stroke	20	4.439	7.060	6.271
30.	p/n	monkey	15	3.860	8.169	6.247

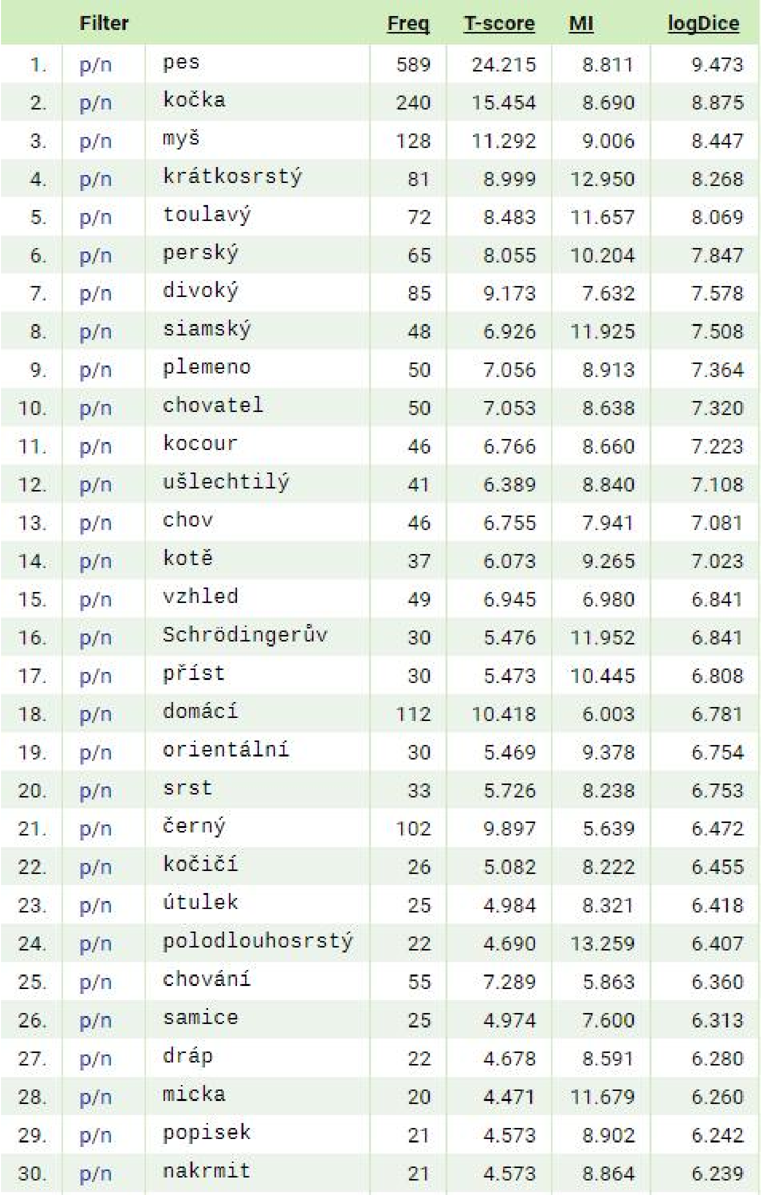
word or lemma collocate?

word (logDice, -3 +3)



	Filter		Freq	T-score	MI	logDice
1.	p/n	psy	124	11.123	9.798	8.603
2.	p/n	psů	101	10.038	9.692	8.341
3.	p/n	pes	108	10.361	8.375	8.077
4.	p/n	psi	80	8.930	9.333	8.001
5.	p/n	psa	83	9.076	8.067	7.722
6.	p/n	kočka	67	8.167	8.780	7.689
7.	p/n	myš	55	7.407	9.652	7.577
8.	p/n	myší	55	7.405	9.418	7.551
9.	p/n	kočky	59	7.660	8.503	7.486
10.	p/n	Vzhled	41	6.400	11.239	7.269
11.	p/n	domácí	85	9.121	6.543	7.019
12.	p/n	krátkosrsté	33	5.744	13.353	6.989
13.	p/n	divoká	31	5.559	9.379	6.798
14.	p/n	chov	30	5.464	8.710	6.695
15.	p/n	černá	33	5.712	7.455	6.600
16.	p/n	psům	25	4.996	10.191	6.545
17.	p/n	vaše	56	7.373	6.089	6.505
18.	p/n	POPISEK	24	4.895	10.341	6.493
19.	p/n	Kočka	24	4.893	9.716	6.470
20.	p/n	toulavé	22	4.690	12.353	6.402
21.	p/n	koťata	22	4.687	10.478	6.375
22.	p/n	kočku	24	4.883	8.245	6.357
23.	p/n	Kočky	22	4.685	9.880	6.356
24.	p/n	krátkosrstá	21	4.582	12.492	6.336
25.	p/n	perské	21	4.581	11.297	6.325
26.	p/n	divoké	24	4.878	7.842	6.303
27.	p/n	micky	19	4.359	13.960	6.198
28.	p/n	chování	45	6.583	5.739	6.170
29.	p/n	mývalí	18	4.242	12.478	6.115
30.	p/n	U	76	8.487	5.239	6.066

lemma (logDice, -3 +3)



	Filter		Freq	T-score	MI	logDice
1.	p/n	pes	589	24.215	8.811	9.473
2.	p/n	kočka	240	15.454	8.690	8.875
3.	p/n	myš	128	11.292	9.006	8.447
4.	p/n	krátkosrstý	81	8.999	12.950	8.268
5.	p/n	toulavý	72	8.483	11.657	8.069
6.	p/n	perský	65	8.055	10.204	7.847
7.	p/n	divoký	85	9.173	7.632	7.578
8.	p/n	siamský	48	6.926	11.925	7.508
9.	p/n	plemeno	50	7.056	8.913	7.364
10.	p/n	chovatel	50	7.053	8.638	7.320
11.	p/n	kocour	46	6.766	8.660	7.223
12.	p/n	ušlechtilý	41	6.389	8.840	7.108
13.	p/n	chov	46	6.755	7.941	7.081
14.	p/n	kotě	37	6.073	9.265	7.023
15.	p/n	vzhled	49	6.945	6.980	6.841
16.	p/n	Schrödingerův	30	5.476	11.952	6.841
17.	p/n	příst	30	5.473	10.445	6.808
18.	p/n	domácí	112	10.418	6.003	6.781
19.	p/n	orientální	30	5.469	9.378	6.754
20.	p/n	srst	33	5.726	8.238	6.753
21.	p/n	černý	102	9.897	5.639	6.472
22.	p/n	kočičí	26	5.082	8.222	6.455
23.	p/n	útulek	25	4.984	8.321	6.418
24.	p/n	polodlouhosrstý	22	4.690	13.259	6.407
25.	p/n	chování	55	7.289	5.863	6.360
26.	p/n	samice	25	4.974	7.600	6.313
27.	p/n	dráp	22	4.678	8.591	6.280
28.	p/n	micka	20	4.471	11.679	6.260
29.	p/n	popisek	21	4.573	8.902	6.242
30.	p/n	nakrmit	21	4.573	8.864	6.239

Semantic preference and prosody

Semantic prosody

- Stubbs (2002): *„there always semantic relations between node and collocates, and among the collocates themselves“*

- semantic prosody = the collocational meaning arising from the interaction between a given node word and its collocates
  - primary function: to express speaker/writer attitude or evaluation
  - semantic prosodies are typically negative (Sinclair: *happen, set in*)
  - semantic prosody operates beyond the meanings of individual words (*personal, price* v. *personal price*)

  - negative: *cause, commit, end up –ing, signs of, underage, teenager, sit through, bordering on, a recipe for*
  - positive: *provide, career*

Semantic preference

- Stubbs (2002)
- semantic preference = the meaning arising from the common semantic features of the collocates of a given node word
  - defined by a lexical set of frequently occurring collocates sharing some semantic features
  - e.g. *large* – typically collocates with items from the same semantic set indicating ‚quantities and sizes'

- s. preference and s. prosody are two disctint yet interdependent collocational meanings with different operating scopes:
  - semantic preference: feature of the collocates, relates the node item to another item from a particular semantic set
  - semantic prosody: feature of the node word, can affect wider stretches of text

Collocation dictionaries

- The BBI Combinatory Dictionary of English
  - first published in 1986 (revised ed. 1997)
  - many sources were used, incl. internet, the BNC, Quirk's Grammar...
  - 14 000 entries, 70 000 collocations
  - collocations are listed under the noun

- Oxford Collocation Dictionary
  - includes the most frequent words
- MacMillan Collocation Dictionary
  - Rundell: omits the most frequent words as their collocates are usually well-known and they are freely combinable (?)

Thank you for your attention!

Questions?

CZECH NATIONAL CORPUS

SEMINAR

Reading

common reading:

Lindquist, H. (2011). Looking for lexis. In *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press, pp 51-57.

Alsina, V. & DeCesaris, J. (2002). Bilingual lexicography, overlapping polysemy, and corpus use. In Bengt Altenberg & Sylviane Granger, *Lexis in Contrast.* Amsterdam/Philadelphia: John Benjamins, pp. 215-229.

Discussion

- What does a corpus lexicographer do to extract a meaning of a word from a corpus?
- How is a dictionary headword usually organized?
- How can the individual meanings of a word (or senses) be ordered in a dictionary?
- What belongs and what does not belong to a collocation dictionary?
- What is semantic prosody and can you think of an example in your mother tongue?
- How can monolingual dictionaries be useful in bilingual lexicography?