Breeds of Cooccurrence

Dialogue 2013, Bekasovo

Denis Paperno
Denis Khachko
Anna Roytberg
Mikhail Roytberg

Abstract

- We propose a classification of statistical collocations by nature of cooccurrence : phrases, repeats, clustering
- Develop criteria for their identification
- Illustrate application of those criteria to the Brown Corpus

Notion of collocation

- (Statistical) collocation: pair of words that tends to cooccur
- Special case: lexicographic collocation, conventional phrases that express complex concepts, e.g. *strong tea*
- Applications of collocation:
  - Lexicography
  - WSD
  - Cooccurrence as basis of DSMs etc….

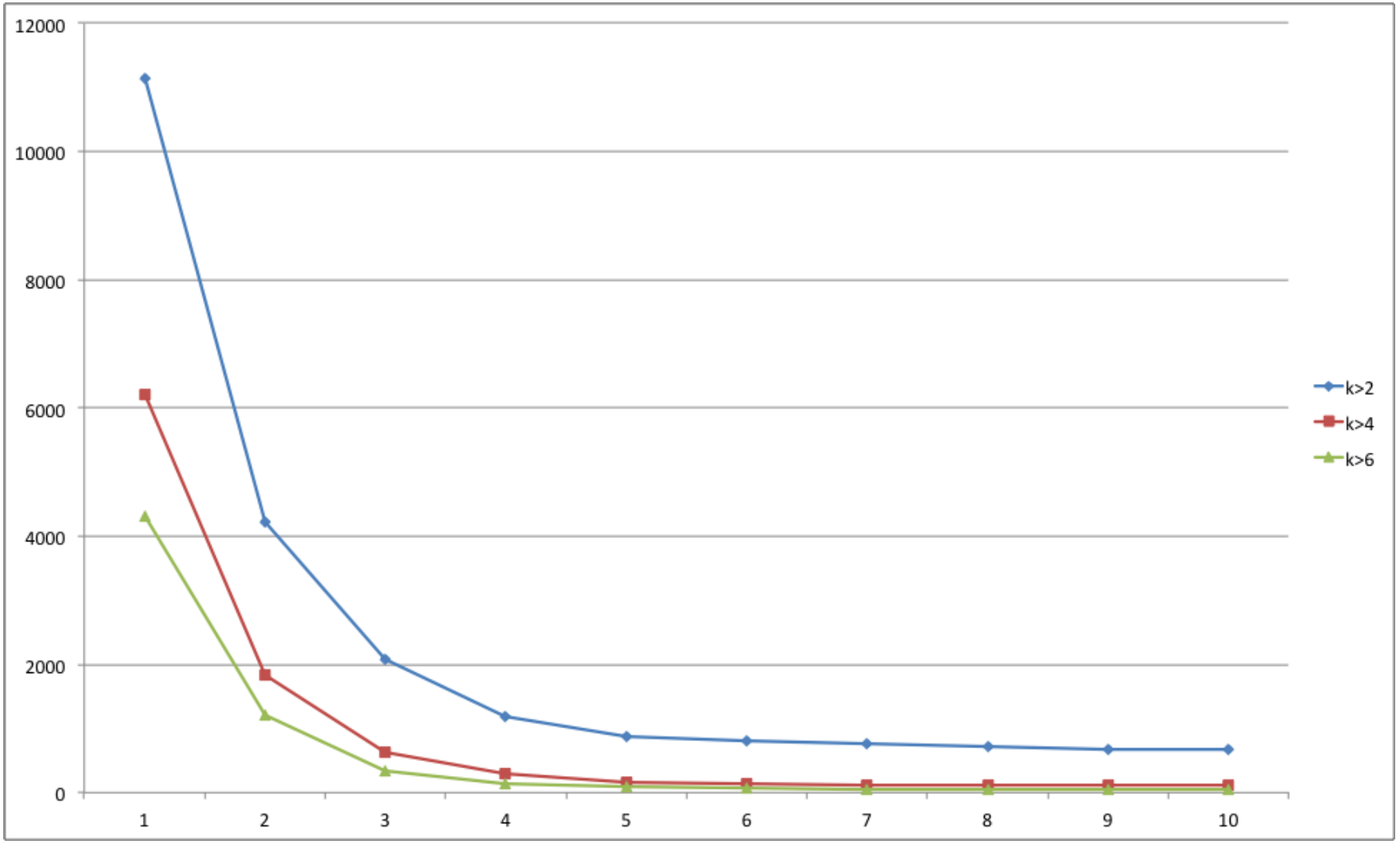
Illustrative Material: Brown Corpus

- Brown corpus (ca. 1 million words)
- Small, well balanced, carefully constructed
- From nltk.googlecode.com [Bird et al. 2009]
- Retagged and lemmatized using Freeling [Carreras et al. 2004]

Identification of collocations

- Cooccurrence criterion: z-score
- $Z = \frac{x-\mu}{\sigma}$ where x is the observed number of pairs, $\mu \approx \sigma^2$ is the expected one; Z=(O-E)/√E
- First word in pair is a content word
- What counts as frequent: *Z*>5, count>4 (2,6)
- We analyze pairs at particular distances (1-10)

Distribution of collocates



Classification

- - «phrases» are multiword expressions with an immediate syntactic relation between words;
- - «repeats» are members of (nearly) identical text fragments such as legal formulae;
- - «clustering» collocations conditioned by corpus heterogeneity;
- Conjecture: no substantial class of word association beyond these three.

Phrases

- syntactically related associated words, in particular, idioms and lexicographic collocations, e.g. *strong tea*
- Lexicographic collocations are perhaps the best studied type of cooccurrence
- **Feature:** tend to occur at short distances
- Typically, within a window ±4 words
- We can exclude lexicographic collocations by considering cooccurrence at greater distances

Examples

N	Word 1	Word 2	Distance	Frequency	Z score
1	real	estate	1	24	231.30
2	urethane	foam	1	12	374.66
3	arc	voltage	2	5	160.49
4	great	deal	2	43	138.93
5	play	role	3	10	45.45
6	write	letter	3	12	33.03

Repeats

- **Repeats** are long phrases repeated as wholes, exactly or with some variation
- **Feature** of word pairs in repeats: other words from the same template are aligned with these and tend to cooccur
- Repeats can be thought of as very long collocations or idioms

Example of a repeat

- *In testimony whereof I have hereunto set my hand and caused the seal of the State to be affixed this 17th day of May in the year of Our Lord one thousand nine hundred and sixty-one.*
- 7 occurrences in one document of the Brown corpus, inflating association scores between words in the template

Identifying repeats

- Criterion: entropy of context

$$E=-\sum_i (P(i)*\ln(P(i)))$$

where i ranges over words, and P(i) is the probability of word i in the given position.

- Entropy averaged across positions
- Heuristic cutoff point: average entropy of context >0.8

Entropy of context: *make sure*

_Position	Occurrence 1	Occurrence 2	Occurrence 3	Occurrence 4	Occurrence 5	entropy
-10	For	and	contract	give	one	1.60944
-9	shooting	any	and	you	on	1.60944
-8	the	other	find	a	the	1.33218
-7	interiors	champion	out	hand	right	1.60944
-6	of	of	He	We	If	1.33218
-5	the	justice	could	straightened	they	1.60944
-4	famous	that	conceivably	Pops	are	1.60944
-3	ante-bellum	he	have	up	Japs	1.60944
-2	Southern	needs	wished	and	Let	1.60944
-1	mansions	to	to	I	us	1.33218
0	make	make	make	make	make	0
1	sure	sure	sure	sure	sure	0
2	your	not	Rev	there	first	1.60944
…	…	…	…	…	…	…

Clustering

- Corpora may happen to contain some text(s) with a dense concentration of two words.
- **Feature**: Within the “dense” subcorpus, there may be no correlation between the two words
- In the corpus as a whole, the two words tend to cooccur to the extent both are associated with the “dense” subcorpus

Identifying Clustering

- Expected occurrence calculated based on a probabilistic model where word probabilities are estimated for each document

$$E=\Sigma_D \left(f_{1(D)}*\left(f_{2(D)}/ N_D \right)\right)$$

where D ranges over all documents, ND is the size of D, f1(D), f2(D) are the frequencies of w1, w2 in D. This calculation is valid regardless of how diverse document sizes are.

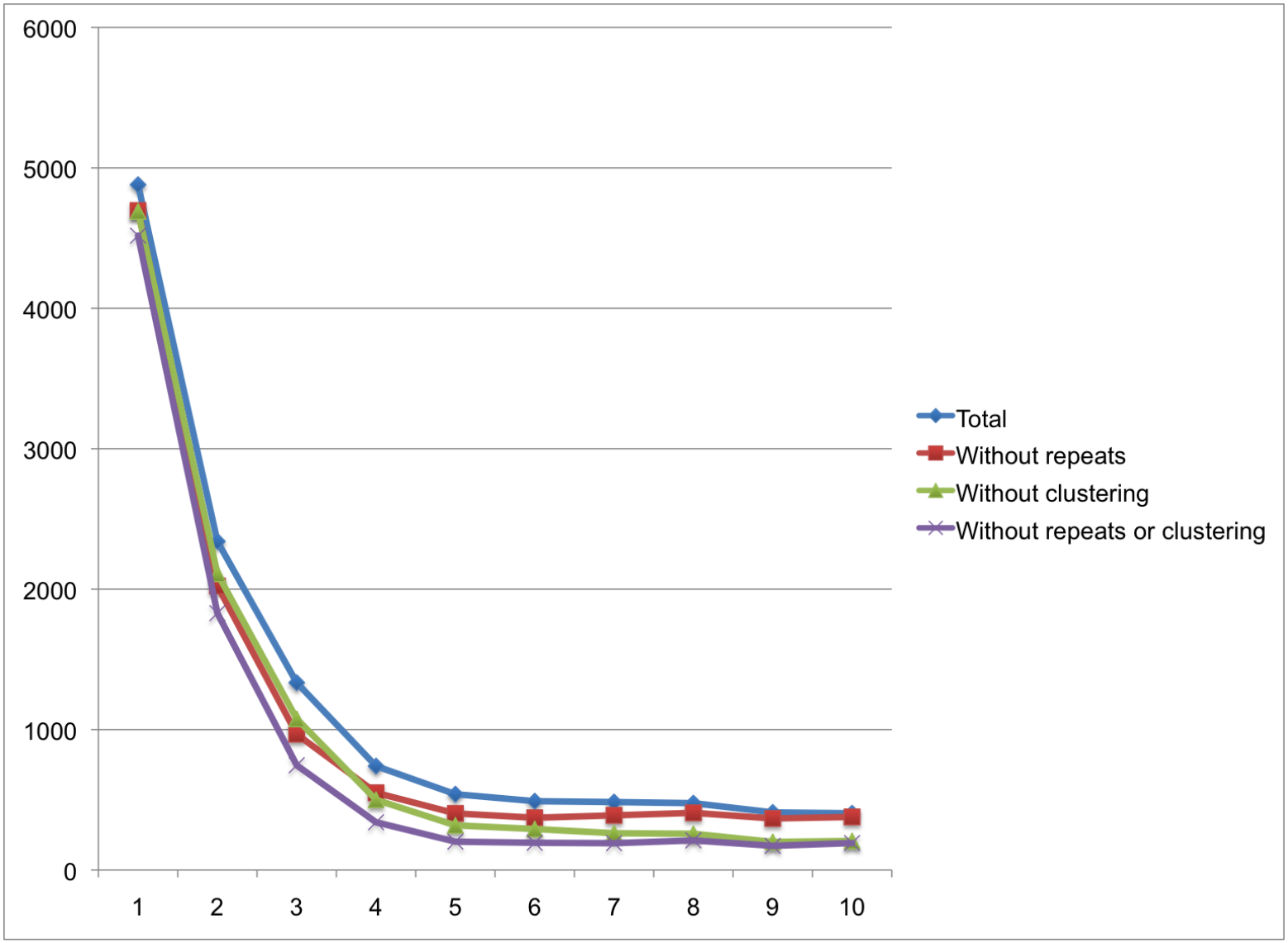
- "Corrected" score calculated based

Examples of Clustering

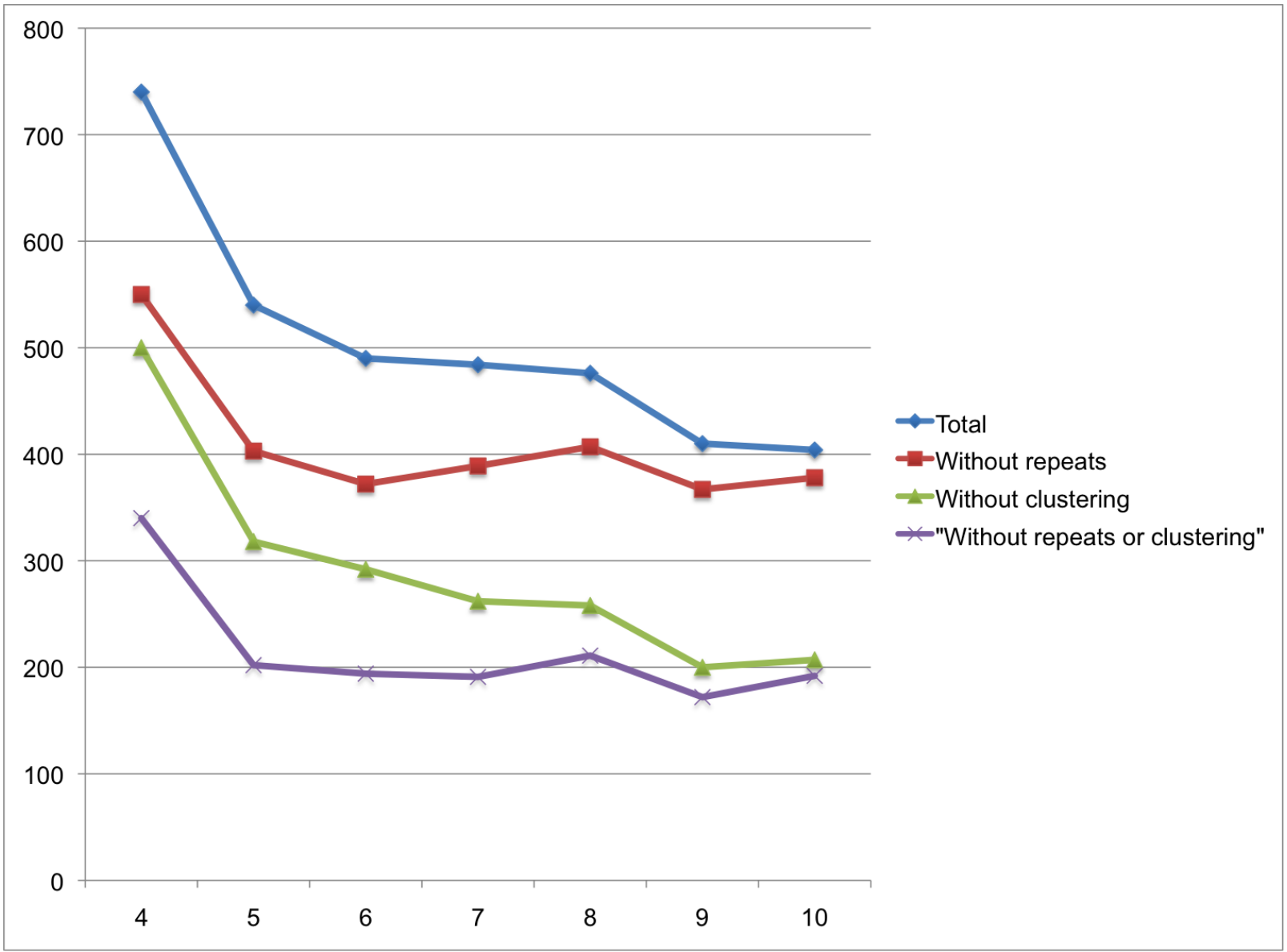
N	Word 1	Word 2	Z	Corrected Z
1	member	church	14.35	4.19
2	student	college	18.05	3.26
3	state	federal	15.73	4.44

- Clustering pairs reveal an association through topic (or genre, or style); this includes sameness of semantic field

Distribution of different types



Distance 4 and up



Is the Classification Exhaustive?

- distances (greater than 5) the majority of pairs belong to the “clustering” collocations and repeats.
- Most of the rest is noise, e.g. pairs of words which exhibit no meaningful relation (*say – New York*, *number – eye, so – work*).
- Raising the frequency threshold essentially eliminates those “remote” collocations.

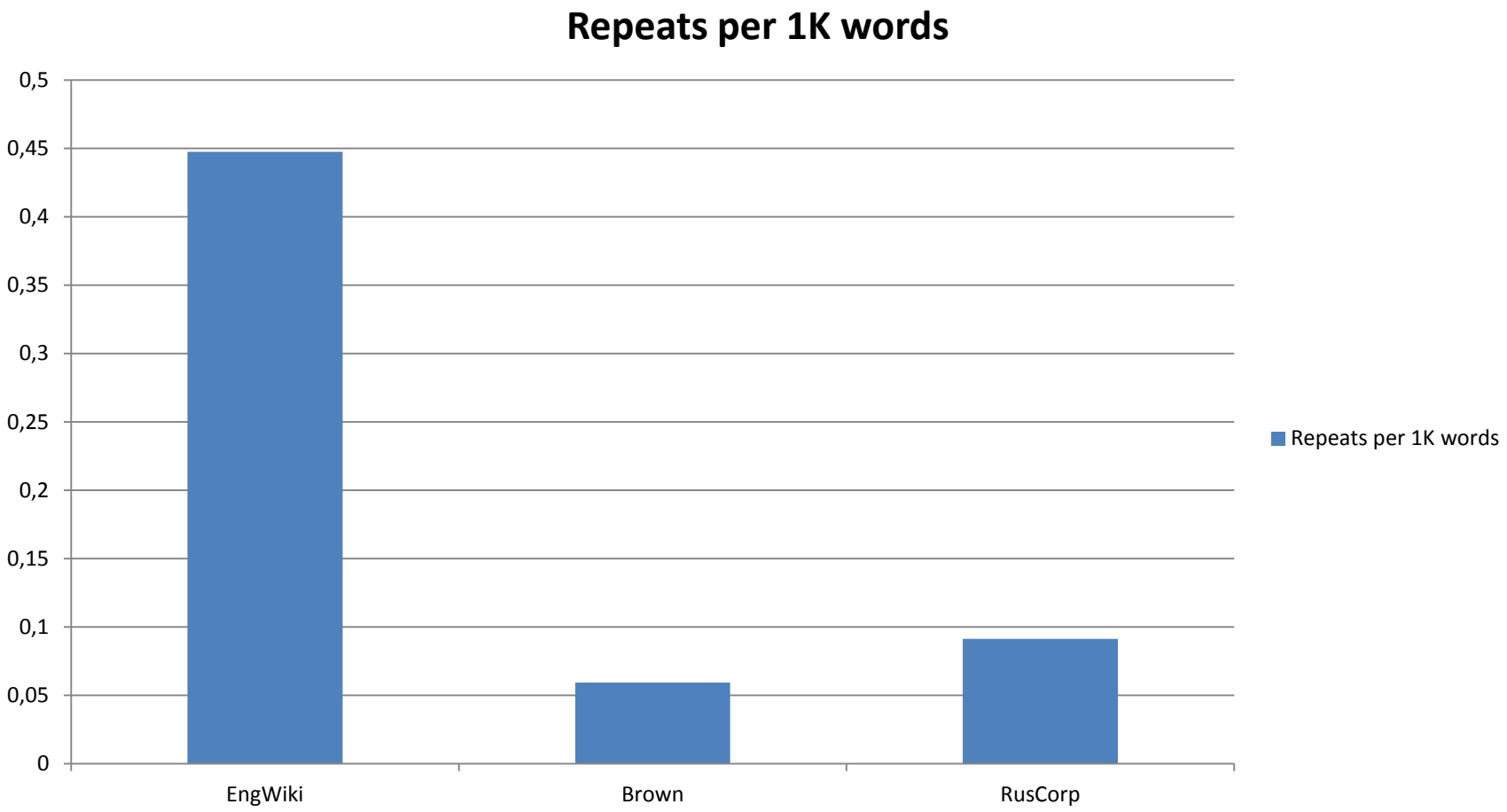
Another theoretically possible type

- Long-range syntactic / discourse link
- E.g. *although…still, on the one hand...on the other hand*
- Brown: almost no examples of this kind. The only exceptions are pairs of markers *first...then, (not) only... (but) also*

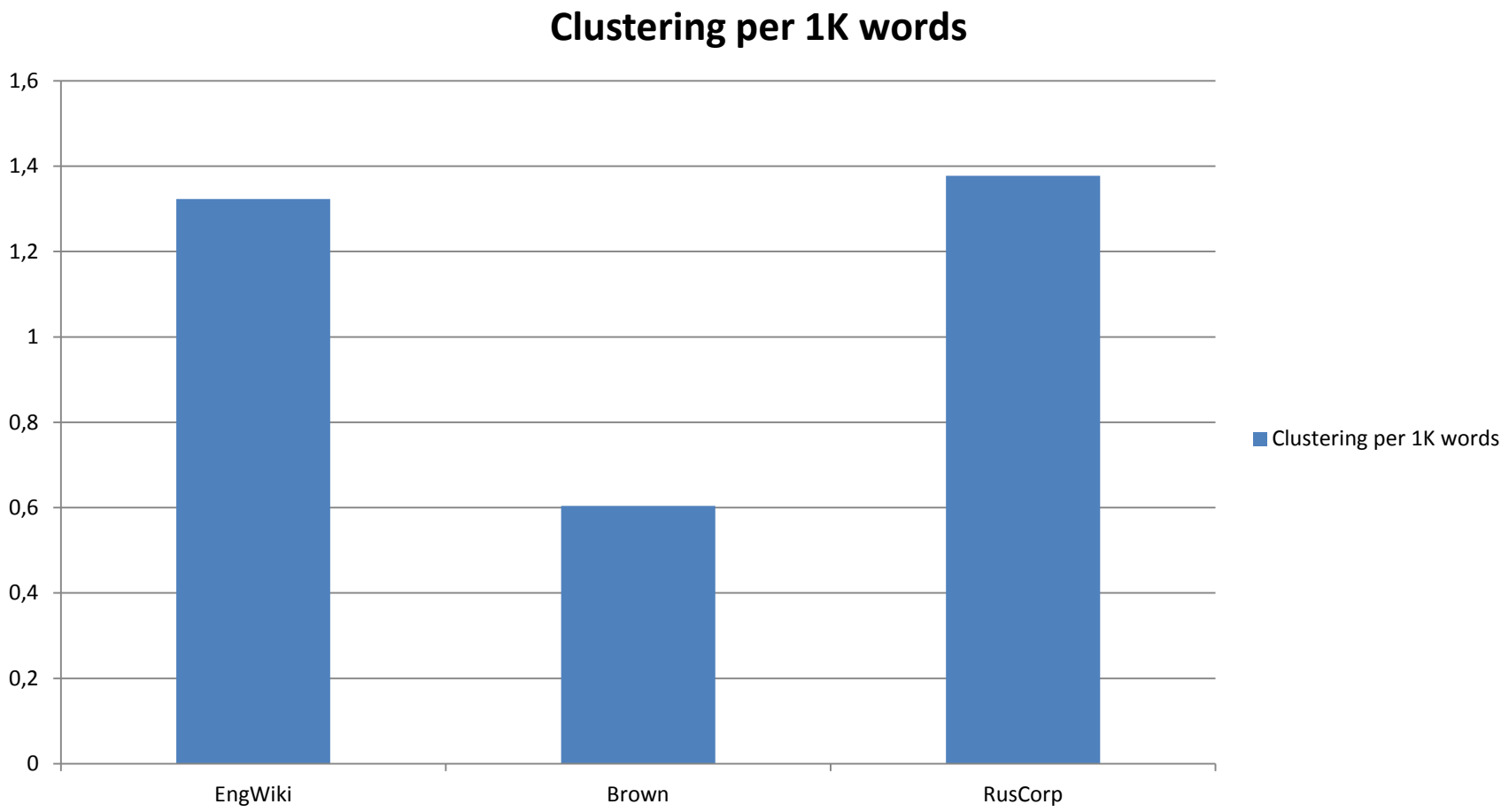
Further directions: corpus evaluation

- abundance of topic-related (clustering) collocates points to a diverse and big corpus
- Abundance of repetition-based collocates points to a corpus that is insufficiently diverse and/or contains a lot of repetition
- Brown (1 mln), disambiguated NCRL (5.8 mln), fragment of English Wikipedia (6.5 mln)

Corpus comparison: repeats at dist.10



Non-repeats at dist.10≈Clustering



Conclusions

- We propose a classification of statistical collocations by nature of cooccurrence : phrases, repeats, clustering
- Develop criteria for their identification
- Illustrate application of those criteria to the Brown Corpus
- Our approach can be applied for characterization of linguistic corpora

Thank you!

Denis Paperno	denis.paperno@unitn.it
Denis Khachko	mordol@gmail.com
Anna Roytberg	cvi@yandex.ru
Mikhail Roytberg	mrojtberg@lpm.org.ru

References

1	Francis W. N., Kucera H. (1964), Department of Linguistics, Brown University, Providence, Rhode Island, USA. http://icame.uib.no/brown/bcm.html
2	Tamir R., Rapp R.(2003), Mining the Web to Discover the Meanings of an Ambiguous Word. IEEE International Conference on Data Mining - ICDM , pp. 645-648.
3	Bell E.J.L.(2007), Collocation Statistical Analysis Tool: An evaluation of the effectiveness of extracting domain phrases via collocation. B.Sc. Dissertation, Lancaster University
4	Evert S. (2004), The Statistics of Word Cooccurrences: Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgar
5	Fillmore Ch., Kay P., O'Connor M. (1988), Regularity and idiomaticity in grammatical constructions: The case of LET ALONE, Language, Vol. 64, 501-518.
6	Firth J.R.(1957) Modes of Meaning, Papers in Linguistics 1934-51, pp. 190-215, Oxford University Press.
7	Halliday M.A.K. (1961), Categories of the Theory of Grammar, Word 17, 241-92.
8	Herbst T. (1996) What Are Collocations: Sandy Beaches or False Teeth? English Studies No. 4, pp 379-393.
9	Manning C., Schutze H.(1999) Foundations of Statistical Natural Language Processing, MIT Press. Cambridge.
10	Mel'cuk, I. (1998). Collocations and Lexical Functions. Cowie, A.P. (ed.), Phraseology. Theory, Practice and Applications, Oxford University Press, 23-53 Oxford.
11	Mitrofanova O.A., Belik V.V. , Kadina V.V., 2008, Corpus Analysis of Selectional Preferences of Frequent words in Russian [Korpusnoe issledovanie sochetaemostnyh predpochtenij chastotnyh leksem russkogo jazyka], Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2008». Moscow.
12	Nesselhauf N.(2004) Collocations in a Learner Corpus, Amsterdam/Philadelphia, Benjamins.
13	Padró L., Stanilovsky E. (2012) FreeLing 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey.
14	Sinclair J. (1991), Corpus, Concordance, Collocation, Oxford University Press, Oxford.
15	Sinclair, J., Carter, R. (2004) Trust the Text. Language, Corpus and Discourse, Routledge, London/New York.
16	Zaharov V.P., Hohlova M.V. (2010) Study of Effectiveness of Statistical Measures for Collocation Extraction on Russian Texts, Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Dialog`2010