A Comparison of Co-occurrence and Similarity Measures as Simulations of Context

Stefan Bordag

Natural Language Processing Department, University of Leipzig
sbordag@informatik.uni-leipzig.de

Abstract. Observations of word co-occurrences and similarity computations are often used as a straightforward way to represent the global contexts of words and achieve a simulation of semantic word similarity for applications such as word or document clustering and collocation extraction. Despite the simplicity of the underlying model, it is necessary to select a proper significance, a similarity measure and a similarity computation algorithm. However, it is often unclear how the measures are related and additionally often dimensionality reduction is applied to enable the efficient computation of the word similarity. This work presents a linear time complexity approximative algorithm for computing word similarity without any dimensionality reduction. It then introduces a large-scale evaluation based on two languages and two knowledge sources and discusses the underlying reasons for the relative performance of each measure.

1 Introduction

One way to simulate associative and semantic relations between words is to view each word as a distinct entity. That entity may occur in a linear stream of sentences or other easily observable linguistic units. It is then possible to measure the statistical correlation between the common co-occurrence of such entities (i.e. words) within these units [1, 2]. If additional knowledge such as word classes or morphological relatedness is available, this model allows to construct a variety of applications that depend on knowledge about word relatedness, but do not necessarily need this knowledge to be precise. For example, it is sufficient to know the most significant co-occurring word pairs in a corpus to enable the creation of a helpful tool for extraction of collocations, idioms or multi-word-expressions [3–5]. Similarly, knowledge about contextual similarity modeled as co-occurrence vector comparisons helps to build thesaurus construction tools such as the Sketch engine [6] or to design specific semi-automatic algorithms that create approximations of a thesaurus [7–10].

Assuming the simple vector-space model where each word defines a new dimension, the question arises how exactly significant co-occurrence or word similarity is to be modeled. Several variations of the same underlying vector space model were proposed. One is to apply Latent Semantic Indexing (LSI) to the matrix containing the raw co-occurrence counts of words [11]. However, it is

unclear what the abstract concepts LSI generates are and whether similarity based on raw frequency counts can achieve the best performance in subsequent applications. Additionally, the gain in computational efficiency by operating on the reduced set of dimensions can be easily outperformed by an approximative algorithm that makes use of the fact that the co-occurrence matrix is sparse, see Section 2.3 below.

Another possibility to model the word space is to apply a significance measure to the co-occurrence counts. Then the ranking of globally most significant word pairs can be used directly as an indicator for possible idiomatic usage of these words [5]. Alternatively, the local ranking of most significant co-occurrences of any word can be used as a condensed contextual representation of that word. These contexts allow to simulate word similarity because obviously, if two words share many significant co-occurrences, then they are probably related to each other. This, in turn, allows to compute rankings of most similar words. These per-word rankings of most significant co-occurrences or most similar words are directly used in applications such as the Sketch engine, automatic thesaurus construction algorithms [12], or document clustering algorithms which accumulate the context representations of the words in a document into a single large vector.

One crucial aspect for all models is the usage of significance and similarity measures. Due to the large number of publications devoted to the development of new measures [13–17], several publications have recently appeared that attempt to measure and compare the performance of some of these measures. However, they either compare only word similarity measures on small evaluations sets (i.e. only 300 nouns [18] or the 80 TOEFL test questions [19]) or are concerned exclusively with the global view [5] which is incompatible with the majority of applications mentioned above. Alternatively, some evaluations are based on creating and then measuring retrieval quality of artificial synonyms [12], a pseudo-disambiguation task [20, 21] or comparing the output of the measures to thesauri [22] or a combination of several of several such methods [23, 24, 2].

Contrary to the previous evaluation efforts this work provides a robust, large-scale evaluation of the most common measures and a discussion of the reasons for the observed differences based on proper statistical significance tests (i.e. not the t-test) to gauge the observed differences.

2 Measures

The measures to be evaluated are inherently divided into two consecutive steps. In order to compare words for similarity, the first operation is to observe co-occurrence frequencies n_{AB} between any word A and B. These are then interpreted by a co-occurrence significance measure, given the individual word frequencies n_A, n_B and the corpus size n. Out of the many possible measures those were chosen that are either frequently used in related work or are statistically well-founded.

2.1 Co-occurrence Measures

As most measures have already been described in great detail, the derivation of most measures is only referenced to in this work. However, for some co-occurrence measures their explicit form with respect to the four observable variables n_{AB}, n_A, n_B and n is given, which in some cases helps to avoid difficulties with too small probability values or with interpretation ambiguities.

The **baseline** for (sentence-wide) co-occurrence significance is to assume that higher frequency means higher significance. In order to relativize a high co-occurrence frequency of A with B with the individual frequencies of the two words, it is possible to compute the **Dice coefficient** [25] as an interpretation of the observed variables.

Assuming that the probability of the occurrence of a word in a sentence $p(A)$ can be approximated by the expression n_A/n then the probability of the co-occurrence of two words A and B in a random corpus should equal $p(A) \cdot p(B) = \frac{n_A \cdot n_B}{n^2}$. In the **Mutual Information** measure [13] this probability is compared to the conditional probability $p(A, B) = \frac{n_{AB}}{n}$ that can be derived from the observed data:

$$sig_{MI}(A, B) = \log_2 \frac{n \cdot n_{AB}}{n_A \cdot n_B} \tag{1}$$

Aware of the problems with MI, especially regarding its preference of low-frequent words, lexicographers modified it (**Lexicographers Mutual Information**) by an additional multiplication with the co-occurrence frequency [6]:

$$sig_{LMI}(A, B) = n_{AB} \log_2 \frac{n \cdot n_{AB}}{n_A \cdot n_B} \tag{2}$$

The **log-likelihood test** [14] uses the generalized likelihood ratio λ to compare two parametrized (binomial in this case) distributions with each other. The first set consists of parameters as expected from the independence assumption. The second derives it's parameters from the observed frequencies. Taking $-2log\lambda$ of the ratio, i.e. the probability of the the observed values, transforms it into a significance value, which is χ^2 distributed, so that the respective thresholds can be used. For example, one degree of freedom and a confidence level of 0.025 means that any value above 5.02 is significant with an error probability of 2.5%. To avoid problems with numerically too extreme probability values [5], it is possible to use the following equivalent and explicit but lengthy form that represents the ratio λ:

$$\lambda = \begin{bmatrix} n \log n - n_A \log n_A - n_B \log n_B + n_{AB} \log n_{AB} \\ + (n - n_A - n_B + n_{AB}) \cdot \log (n - n_A - n_B + n_{AB}) \\ + (n_A - n_{AB}) \log (n_A - n_{AB}) + (n_B - n_{AB}) \log (n_B - n_{AB}) \\ - (n - n_A) \log (n - n_A) - (n - n_B) \log (n - n_B) \end{bmatrix} \tag{3}$$

The significance is then computed as follows:

$$sig(A,B)_{lgl} = -2\log\lambda \tag{4}$$

This test is only one-sided, in that it does not distinguish significant co-occurrence from significant non-co-occurrence. To amend this, a second significance can be defined:

$$sig(A,B)_{lgl2} = \begin{array}{c} -2\log\lambda \text{ if } n_{AB} < \frac{n_A \cdot n_B}{n} \\ 2\log\lambda \text{ else} \end{array} \tag{5}$$

If the frequency of most words is much smaller than the corpus size, the Poisson distribution is a good approximation of the binomial distribution, which leads to the **Poisson significance measure** [16]. Using it instead of the binomial distribution results in a formula that contains the term $\ln n_{AB}!$ which is hard to handle numerically for larger n_{AB}. However, in such cases it is possible to use approximations, such as Stirling's formula, which (with $\lambda = \frac{n_A \cdot n_B}{n}$) results in the following explicit form:

$$sig_{ps1}(A,B) \approx n_{AB}\,(\ln n_{AB} - \ln\lambda - 1) + \tfrac{1}{2}\ln 2\pi n_{AB} + \lambda \tag{6}$$

Another presumably acceptable [17] approximation $\ln k! = k\ln k - k + 1$ which can be further simplified to $\ln k! = k\ln k$, results in the following significance measure:

$$sig_{ps2}(A,B) \approx n_{AB}\,(\ln n_{AB} - \ln\lambda - 1) \tag{7}$$

However, this simplification introduces a systematic error which results in an increasing positive discrepancy for larger n_{AB}. The effect is that this approximation systematically overrates larger co-occurrence frequencies over small ones, which also explains the varying performance in the evaluations below.

The **z-score** and the **t-score** (from the t-test) are two commonly used measures [5]. The z-score divides the difference between the expected and the observed value by the expected value $\frac{n_A \cdot n_B}{n^2}$:

$$sig(A,B)_{z-sc} = \frac{n_{AB} - \frac{n_A \cdot n_B}{n^2}}{\sqrt{\frac{n_A \cdot n_B}{n^2}}} \tag{8}$$

The t-score, applied to this task according to the 'standard way' [26] differs from the z-score only in dividing by the observed value n_{AB} instead of the expected one:

$$sig(A,B)_{t-sc} = \frac{n_{AB} - \frac{n_A \cdot n_B}{n^2}}{\sqrt{n_{AB}}} \tag{9}$$

Except for the Dice coefficient most measures produce values which are comparable to each other only if one of the two words is the same (i.e. in the local ranking case). In particular, it does not make sense to compare $sig(A,B)_{lgl} = 50$ with $sig(C,D)_{lgl} = 10$, because even though the 50 is numerically larger, the corresponding word frequencies, for example $n_A = 1000$ and $n_B = 1000$, might make it less important than the 10 with $n_A = 10$ and $n_B = 10$.

2.2 Similarity Measures

The similarity measures included in the evaluation are standard measures such as the cosine or euclidian distance and can be applied in different ways. It is possible to ignore the significance values computed by the significance measures above by transforming all vectors into binary vectors and then computing the cosine, for example. Alternatively, it is possible to keep the values and use the same measure. The following listing gives the similarity measures used to compare two vectors in the evaluation below (a more detailed listing including formulae can be found in related word, for example [27]):

- baseline (base) - the number of matching non-zero elements in both vectors
- overlap (over) - baseline divided by the minimum of non-zero elements in both vectors
- Dice (Dice) - baseline multiplied by 2 and divided by the amount of non-zero elements in both vectors
- binary cosine (cbin) - angle between the binary versions of both vectors
- cosine (cos) - angle between both vectors
- city block metric (L1) - the sum of the pairwise absolute differences between each value of both vectors
- euclidian distance (L2) - the square root of the pairwise squared differences between each value of both vectors
- Jensen-Shannon divergence (JS) - transforms both vectors into probability distributions, builds a mean distribution and measures the mean divergence of both original distributions to the mean distribution

Note that the JS divergence is defined only when applied on the plain frequency counts of co-occurrence, not on interpreted values. Technically however, any vector can be transformed into a probability distribution, but other than for frequency counts, the result is meaningless.

2.3 Computing Similarity

In order to compute the similarity between all words it is not necessary to compare each word with each other or to reduce the dimensionality of the entire vector space. Obviously, only words that are co-occurrences of co-occurrences of the input word are candidates to share any co-occurrences with it. Due to the power-law distribution of word frequency, for most words this candidate list is short (on the order of less than 100 words). In any case, but especially for the remaining very frequent words it is possible to restrict the search for candidates to use only the most significant co-occurrences to find new candidates.

This results in an algorithm with two approximation parameters which make the complexity of the entire algorithm linear, instead of quadratic. This is combined with programming the vector representations in a way that a vector uses only as much memory as there are non-zero entries in it. Additionally to the

linear time-complexity of the algorithm the resulting constant memory requirement also enables computing similarity on arbitrarily large corpora without the need for costly hardware.

In the experiments reported here each word was compared with a maximum of 10 000 other words. Additionally, the maximal amount of values to be compared ordered by decreasing significance was restricted to 200. These approximations affected only 2% of all comparisons but decreased run-time significantly. However, these approximations skew the binary vector similarity measures: Since especially the baseline counts only matching non-zero values it should produce the exact same results irrespective of the underlying co-occurrence significance measure. Yet, since some values in the vectors are ignored depending on the particular co-occurrence measure, the performance varies, as shown below.

3 Evaluation

The experimental setup comprises a corpus and a gold standard and is repeated for two languages, English and German. The gold standards used are, respectively, WordNet [28] and GermaNet [29].

First, for each co-occurrence measure all sentence-wide co-occurrences (including insignificant ones) on the raw occurrences of words without any POS-tagging or other preprocessing are computed (apart from basic tokenization and sentence splitting). Then, for each word and for each measure the ranking of most similar words is obtained, which makes a total of $11 + 11 \cdot 8$ result sets. Then each result set is cut to contain only words that are in the gold standard and among the 100 000 most frequent words. Additionally, the ranking for each remaining word is cut to contain 100 words. The experiment was repeated for the BNC with WordNet and for a German subcorpus of the 'Wortschatz Projekt' with GermaNet. For English this leaves 35 966 input words with 100 output words each and for the German subcorpus 21 686 words. In the gold standards only those words were counted, which occured at least once in the corresponding corpus, i.e. 57 990 out of 146 212 for WordNet and 40 703 out of 52 620 for GermaNet. On average for each valid input word 59.9 relevant output words are found according to WordNet and 33.3 according to GermaNet. In both cases roughly 83% are cohyponyms.

The evaluation consists of evaluating the average quality of the ranking produced for each word. For this purpose, any word that stands in any relation with the input word in the gold standard is counted as relevant, similarly to relevant and irrelevant documents in IR. Both semantic nets were modified so that they also contain the cohyponymy relation (assuming that words sharing a direct hyperonym are cohyponyms). There are some problematic issues concerning this kind of evaluation, most importantly with the unknown upper bounds. On the one hand, an algorithm might compute many correct word pairs, which are all counted as wrong. The smaller the gold standard, the more severe is the effect on the evaluation. On the other hand, the gold standard might contain many annotated word pairs which are not observable in the corpus.

	base	Dice	MI	LMI	t-sc	z-sc	lgl	lgl2	ps1	ps2
only	1.95	*8.35*	6.35	6.82	1.47	*8.08*	*7.79*	*7.79*	*8.13*	*8.26*
base	5.37	7.95	6.09	7.77	4.99	7.14	**8.79**	**8.78**	**8.65**	10.29
over	5.37	7.56	5.90	5.27	4.80	5.89	5.71	5.50	6.27	7.84
Dice	5.43	8.12	6.12	7.89	5.06	7.35	**8.98**	**8.98**	**8.82**	10.37
cos	3.29	8.09	6.30	8.49	5.63	6.90	7.96	8.80	8.82	9.16
cbin	5.44	8.12	6.10	7.93	5.09	7.29	**8.96**	**8.94**	**8.77**	10.17
L1	5.88	3.74	4.41	6.23	5.97	3.67	5.30	5.53	5.32	4.23
L2	5.70	3.84	3.93	7.07	5.93	3.52	6.36	7.05	6.17	6.03
JS	5.68	3.80	3.37	4.81	5.49	3.52	4.21	4.16	3.36	3.54

Table 1. Precision for 5 most significant or similar words in % for BNC measured on WordNet for all measure combinations based on 35 966 test words. Word pairs in any relation are counted as relevant. Two groups of measure combinations that do not differ significantly are emphasized.

Nevertheless, used on such a large scale, this evaluation gives reliable (with respect to statistical significance) relative performances of the measures. The evaluation measures used are **mean average precision** (MAP) and **precision**. Precision is defined as the number of relevant words found in the ranking, divided by the length of the ranking. It is possible to measure only top 5 words, for example, so that the expected performance in a thesaurus creation scenario is reflected more closely. MAP is defined as the sum of inverse ranks divided by the minimum of relevant words and ranking length and thus represents a combination of precision, recall and ranking quality. If 2 out of 40 relevant words were found at ranking positions 3 and 6 and the algorithm returned 100 words, then MAP in percent in this case is $\frac{\frac{1}{3}+\frac{2}{6}}{min(40,100)} \cdot 100 = 1.6\%$.

3.1 Results

Unsuprisingly, the values in Table 1 and 2 differ and are generally very low. Therefore the following discussion is based on the results of Scheffé's test which is an ANOVA post-hoc test to discover possible interactions in a group of means. Unfortunately, this necessary test has not yet been applied in related discussions. This test produces statements about whether a precision of 8.14% (*ps2_only*) and 7.12% (*MI_only*) differ significantly with an error probability of 0.0097 ($alpha = 0.05$). Usually, the pairwise t-test is applied instead; incorrectly, because the t-test is not able to correctly distinguish groups of possibly related test instances.

When **comparing the performances of co-occurrence significance measures**, Scheffé's test finds three groups of measures. The largest group (group significance 0.247 with $\alpha = 0.05$) includes Dice, z-score, both poisson variants and both log-likelihood variants. The second group is composed of both mutual information variants (group significance 0.807), whereas the remaining

group comprises the t-score and the baseline (group significance 0.664). However, when these co-occurrence frequency interpretations are used to compute word similarity, the similarity computations based on the likelihood (and poisson) co-occurrence measures perform significantly better than any other. Surprisingly, the second (unprecise) poisson approximation *ps2* gives the best results and is found to significantly differ from all other measures. Apparently the systematic overrating of high co-occurrence counts helps, if not taken directly (as in the case of the baseline).

The following example illustrates the differences: The input word A ($n_A = 1000$) in a one million sentences corpus co-occurs 100 times with B ($n_B = 500$) and 150 times with C ($n_C = 2000$). According to the Dice coefficient $sig(A, B)_{Dice} = 0.133$ and $sig(A, C)_{Dice} = 0.1$, which means that the less frequent B is a more significant co-occurrence than C. Contrary to that, the second poisson approximation ($sig(A, B)_{ps2} = 429$ and $sig(A, C)_{ps2} = 497$) results in an inversed ranking. The reason why this has adverse effects on similarity computations is sparsity. The less frequent a word in a co-occurrence vector, the less probable it makes a match with a co-occurrence vector of another word.

The sole difference between the two log-likelihood variants as described above is that the second method differentiates between significant inhibition and attraction. The only effect this seems to have is a slight but statistically insignificant improvement of the subsequent similarity rankings.

As expected, the overrating of infrequent words renders the rankings of the mutual information measure nearly useless compared to the other measures. While the lexicographer's modification indeed helps, it does not help enough to produce significantly better results.

	base	Dice	MI	LMI	t-sc	z-sc	lgl	lgl2	ps1	ps2
only	3.47	*8.59*	7.12	6.90	2.84	*8.48*	*7.71*	*7.71*	*8.06*	*8.14*
base	7.44	10.98	8.78	10.40	6.92	9.18	**12.79**	**12.87**	**12.11**	14.63
over	7.25	10.88	8.77	8.06	6.89	8.67	9.21	9.17	9.71	11.18
Dice	7.55	11.24	8.92	10.59	7.03	9.57	**13.17**	**13.24**	**12.45**	14.70
cos	7.99	11.18	9.26	11.86	8.33	9.05	11.27	11.87	11.36	12.33
cbin	7.59	11.25	8.92	10.68	7.08	9.57	**13.15**	**13.21**	**12.36**	14.28
L1	8.64	5.92	7.02	11.61	9.12	6.54	8.68	8.90	7.85	6.74
L2	8.22	6.46	6.61	10.64	8.75	6.15	9.74	10.23	8.81	8.65
JS	7.98	6.82	5.88	7.69	7.92	6.26	6.49	6.53	5.46	5.19

Table 2. Precision for 5 most significant or similar words in % for the German subcorpus measured on GermaNet for all measure combinations based on 21 686 test words. Word pairs in any relation are counted as relevant. Two groups of measure combinations that do not differ significantly are emphasized.

The baseline - ranking according to co-occurrence frequency - is consistently outperformed by every measure except the t-score throughout both experiments. Hence, 'Yes, we can do better than frequency!', to answer the question formulated earlier [30] on a similar topic.

Comparing the similarity measures results in several surprising observations which are independent on which co-occurrence measure they are based on, except for the baseline and the t-score. First, it seems to be impossible to outperform the baseline. Second, Jensen-Shannon divergence is not the best measure, which apparently contradicts the results of other researchers. And third, the introduced approximations when computing similarity appear to be harmful if the co-occurrence measure has a poor performance.

The baseline disregards the computed significance values and only counts matching non-zero elements in the two vectors to be compared. The other measures either additionally take the amount of non-zero elements into account or weight the non-zero elements according to their relative values. Obviously, such additional information either degradates the results (for the overlap, for example) or does not seem to improve them (Dice). A manual examination of several examples revealed that the reason is a combination of data sparsity and Zipfs Law [31]. As could be expected, given an input word A and a set of words whose co-occurrence vectors have at least one matching element with the co-occurrence vector of A, the amount of matches is power-law distributed. That means that most words have one match, fewer two matches, etc. This also means that especially the words with the highest contextual similarity to A have rapidly falling amounts of matches. The first might have 200 matches, the next only 167, then the next only 150, etc.

As mentioned previously, the Jensen-Shannon divergence is applicable only to frequency counts, because only they can be directly transformed into probabilities. Hence, in the results tables only the results for JS applied on the baseline co-occurrence measure are meaningful. Out of all measures applied on the baseline co-occurrence data, the JS is indeed among the best measures, although not significantly. However, using a proper co-occurrence significance measure such as the second poisson approximation and then computing similarity using for example the Dice measure consistently produces approximately twice as good similarity rankings in both experiments.

In fact, these results suggest that any similarity measure based on pure frequency counts is at a disadvantage against comparing interpreted vectors. While it is unclear, whether for example the generalization to fewer concepts in LSI might alleviate the problem, it might also be the case that information-loss reduces the performance even more. A more direct comparison of LSI to the methods here would be necessary.

Finally, exactly the same observations can be reproduced from the MAP values for both experiments. The only difference is that the MAP values are lower and differ more between the two languages. For example, the *ps2-base* measure combination has a MAP value of 2.58 for English and 4.66 for German. The larger difference between the language is apparently due to the differing

sizes of the knowledge bases which means that for English a lower recall is achieved. WordNet is about twice as large as GermaNet. Hence under similar circumstances it is to be expected that the same algorithm misses twice as much for English.

Additionally to Scheffé's test, Pearson's correlation coefficient helps to quantify the similarities between the various rankings of each measure combination. In the entire matrix of possible measure combinations only two pairs of combinations did not differ: *lg_only* and *lg2_only* had equal results. The only difference between them is that in 4.7% cases the computed significance is negative (resulting in a different ranking). Given that this difference did not affect the retrieval quality at all suggests that only words or word pairs not annotated in the knowledge sources are affected. In fact, only very frequent function words are affected. Because using the *z-sc* and *ps2* significances to compute similarity yields such differing quality of similarity rankings the main differences between these rankings are probably located in a range which is not measurable using WordNet or GermaNet, i.e. again function words.

On average, Pearson's correlation coefficient between all measure combinations ranges from 0.2 to 0.6. The only exceptions achieving values as high as 0.7 or 0.8 are pairs of similarly motivated co-occurrence measures and equal similarity measures such as *lg_base* and *ps1_base* 0.8 or *lg_base* and *lg2_base* 0.96. Such pairs were also found to not differ significantly by Scheffé's test above. It is interesting that some measures that did not perform well such as the *t-sc*, *z-sc* or both Mutual information variants correlate stronger with the log-likelihood measures with values between 0.6 and 0.7 than among each other (for example 0.42 for *t-sc* and *z-sc*). Supporting the finding that using *ps2* for similarity computations significantly outperforms all other co-occurrence measures the Pearson coefficient to the most similar similarity rankings is only 0.79 to *lg_base*. Roughly speaking, according to the results of Scheffé's test, any correlation values below 0.8 entails significant difference with an error probability of less than 0.05.

4 Conclusions

The approximations used in the similarity computation algorithm introduced in this work were shown to have a strong positive effect on the time-complexity of computing similarity. It was also shown that the approximations are only harmful, if the underlying co-occurrence significance measure was ill-chosen.

The interactions between the included measures have been fleshed out and underpinned with sound statistical tests. There are strong indications that any similarity measure dependent on raw frequency counts such as probability distribution divergence measures can be easily outperformed by much simpler comparisons based on co-occurrence significance values instead of frequency counts.

Further research should examine the interactions of measure performance with other factors such as corpus size or word frequency. The evaluation method employed in this work can easily be used to measure the relative performance at computing specific relations. It can be expected, for example, that co-occurrence

rankings contain more syntagmatic relations, whereas similarity rankings should be more paradigmatic (see also [2, 27]. Additionally, a direct comparison of the effects of the various measures on using them for local rankings as in this work or modified versions for global rankings is necessary as well. Especially in order to explore the discrepancy between the reported performance figures for example for the t-score which was found to be the worst measure in this evaluation but one of the best in other evaluations [5] using global rankings.

References

1. Finch, S.P.: Finding Structure in Language. PhD thesis, University of Edinburgh, Edinburgh, Scotland, UK (1993)
2. Sahlgren, M.: The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. PhD thesis, Swedish Intitute of Computer Science, Stockholm, Sweden (2006)
3. Smadja, F.: Retrieving collocations from text: Xtract. Computational Linguistics **19** (1993) 43–177
4. Lin, D.: Extracting collocations from text corpora. In: Proceedings of the First Workshop on Computational Terminology. (1998)
5. Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, University of Stuttgart, Stuttgart, Germany (2004)
6. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D.: The sketch engine. In: Proceedings of Euralex, Lorient, France (2004) 105–116
7. Riloff, E., Shepherd, J.: A corpus-based approach for building semantic lexicons. In Cardie, C., Weischedel, R., eds.: Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP 1997), Somerset, NJ, USA, Association for Computational Linguistics (ACL) (1997) 117–124
8. Roark, B., Charniak, E.: Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In: Proceedings of The 17th International Conference on Computational Linguistics (COLING/ACL), Montreal, Quebec, Canada (1998) 1110–1116
9. Widdows, D.: Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In: Proceedings of the Human Language Technology Conference (HLT) of the NAACL, Edmonton, Canada (2003) 276–283
10. Rohwer, R., Freitag, D.: Towards full automation of lexicon construction. In: Proceedings of Computational Lexical Semantics Workshop at the HLT/NAACL, Boston, MA, USA (2004)
11. Dumais, S.T.: Latent semantic indexing (LSI). In Harman, D.K., ed.: Overview of the Third Text Retrieval Conference (TREC), Gaithersburg, MD, USA, National Institute of Standards and Technology (1995) 219–230
12. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Press, Boston, MA, USA (1994)
13. Church, K.W., Gale, W.A., Hanks, P., Hindle, D.: Using statistics in lexical analysis. In Zernik, U., ed.: Lexical Acquisition: Exploiting On-Line Resources to Build up a Lexicon. Lawrence Erlbaum, Hillsdale, NJ, USA (1991) 115–164
14. Dunning, T.E.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics **19** (1993) 61–74

15. Lee, L.: Measures of distributional similarity. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), College Park, MD, USA (1999) 25–32
16. Holtsberg, A., Willners, C.: Statistics for sentential co-occurrence. In: Working Papers 48. (2001) 135–148
17. Quasthoff, U., Wolff, C.: The poisson collocation measure and its applications. In: Second International Workshop on Computational Approaches to Collocations, Vienna, Austria (2002)
18. Curran, J.R.: From Distributional to Semantic Similarity. PhD thesis, Institute for Communicating and Collaborative Systems, School of Informatics. University of Edinburgh, Edinburgh, Scotland, UK (2003)
19. Terra, E., Clarke, C.L.A.: Frequency estimates for statistical word similarity measures. In: Proceedings of the Human Language Technology Conference (HLT) of the NAACL, Edmonton, Canada (2003) 165–172
20. Gale, W., Church, K.W., Yarowsky, D.: Work on statistical methods for word sense disambiguation. Intelligent Probabilistic Approaches to Natural Language **Fall Symposium Series** (1992) 54–60
21. Schütze, H.: Context space. In: Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, Menlo Park, CA, USA, AAAI Press (1992) 113–120
22. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of The 17th International Conference on Computational Linguistics (COLING/ACL). (1998) 768–774
23. Weeds, J., Weir, D.: Co-occurrence retrieval: A flexible framework for lexical distributional similarity. Computational Linguistics (2005) 439–475
24. Weeds, J.: The reliability of a similarity measure. In: Proceedings of the 5th UK Special Interest Group for Computational Linguistics (CLUK), Manchester, UK (2005)
25. Smadja, F., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: A statistical approach. Computational Linguistics **22** (1996) 1–38
26. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press (1999)
27. Bordag, S.: Elements of Knowledge-free and Unsupervised lexical acquisition. PhD thesis, Department of Natural Language Processing, University of Leipzig, Leipzig, Germany (2007)
28. Fellbaum, C.: A semantic network of English: The mother of all WordNets. Computers and the Humanities **32** (1998) 209–220
29. Hamp, B., Feldweg, H.: GermaNet - a lexical-semantic net for German. In: Proceedings of workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications at the ACL, Madrid, Spain (1997)
30. Krenn, B., Evert, S.: Can we do better than frequency? a case study on extracting pp-verb collocations. In: Proceedings of the Workshop on Collocations at the ACL, Toulouse, France (2001) 39–46
31. Zipf, G.K.: Human Behaviour and the Principle of Least-Effort. Cambridge MA edn. Addison-Wesley (1949)