Normalized (Pointwise) Mutual Information in Collocation Extraction

Gerlof Bouma

Department Linguistik, Universität Potsdam

Abstract. In this paper, we discuss the related information theoretical association measures of mutual information and pointwise mutual information, in the context of collocation extraction. We introduce normalized variants of these measures in order to make them more easily interpretable and at the same time less sensitive to occurrence frequency. We also provide a small empirical study to give more insight into the behaviour of these new measures in a collocation extraction setup.

1 Introduction

In collocation extraction, the task is to identify in a corpus combinations of words that show some idiosyncrasy in their linguistic distribution. This idiosyncrasy may be reduced semantic compositionality, reduced syntactic modifiability or simply a sense that the combination is habitual or even fixed. Typically but not exclusively, this task concentrates on two-part multi-word units and involves comparing the statistical distribution of the combination to the distribution of its constituents through an *association measure*. This measure is used to rank candidates extracted from a corpus and the top ranking candidates are then selected for further consideration as collocations.[1]

There are literally dozens of association measures available and an important part of the existing collocation extraction literature has consisted of finding new and more effective measures. For an extreme example see Pecina (2008a), who in one paper compares 55 different (existing) association measures and in addition several machine learning techniques for collocation extraction. A recent development in the collocation literature is the creation and exploitation of gold standards to evaluate collocation extraction methods – something which is for instance standard practice in information retrieval. Evaluation of a method, say, a certain association measure, involves ranking the data points in the gold standard after this measure. An effective method is then one that ranks the actual collocations in this list above the non-collocations. Four such resources, compiled

[1] In the context of this paper, we will not attempt a more profound definition of the concept of collocation and the related task of collocation extraction. For this we refer the interested reader to Manning and Schütze (1999, Ch. 5) and especially Evert (2007). A comprehensive study of all aspects of collocation extraction with a focus on mathematical properties of association measures and statistical methodology is Evert (2005).

for the shared task of the MWE 2008 workshop, are described in Baldwin (2008), Evert (2008a), Krenn (2008), and Pecina (2008b).

One of the lessons taught by systematic evaluation of association measures against different gold standards is that there is not one association measure that is best in all situations. Rather, different target collocations may be found most effectively with different methods and measures. It is therefore useful to have access to a wide array of association measures coupled with an understanding of their behaviour if we want to do collocation extraction. As Evert (2007, Sect. 6), in discussing the selection of an association measure, points out, choosing the best association measure for the job involves empirical evaluation as well as a theoretical understanding of the measure.

In this paper, we add to the large body of collocation extraction literature by introducing two new association measures, both normalized variants of the commonly used information theoretical measures of *mutual information* and *pointwise mutual information*. The introduction of the normalized variants is motivated by the desire to (a) use association measures whose values have a fixed interpretation; and (b), in the case of pointwise mutual information, reduce a known sensitivity for low frequency data. Since it is important to understand the nature of an association measure, we will discuss some theoretical properties of the new measures and try to gain insight in the relation between them and the original measures through a short empirical study.

The rest of this paper is structured as follows: Section 2 discusses mutual information and pointwise mutual information. We then introduce their normalized variants (Sect. 3). Finally, we present an empirical study of the effectiveness of these normalized variants (Sect. 4).

2 Mutual information

2.1 Definitions

Mutual information (MI) is a measure of the information overlap between two random variables. In this section I will review definitions and properties of MI. A textbook introduction can be found in Cover and Thomas (1991). Readers familiar with the topic may want to skip to Sect. 3.

The MI between random variables X and Y, whose values have marginal probabilities $p(x)$ and $p(y)$, and joint probabilities $p(x, y)$, is defined as:[2]

$$I(X;Y) = \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}. \tag{1}$$

[2] In this paper, I will always use the natural logarithm. Changing the base of the logarithm changes the unit of measurement of information, but this is not relevant in the context of this paper. Further, capital variable names refer to random variables, whereas lowercase ones refer to the values of their capitalized counterparts. Finally, $0 \cdot \ln 0$ is defined to be 0, which means that in a contingency table, cells with zero counts/probability do not contribute to MI, entropy, etc.

The information overlap between X and Y is 0 when the two variables are independent, as $p(x)p(y) = p(x, y)$. When X determines Y, $I(X;Y) = H(Y)$, where $H(Y)$ is the entropy of, or lack of information about, Y, defined as:

$$H(Y) = -\sum_{y} p(y) \ln p(y). \tag{2}$$

When X and Y are perfectly correlated (they determine each other), $I(X;Y)$ reaches its maximum of $H(X) = H(Y) = H(X, Y)$, where $H(X, Y)$ is the joint entropy of X and Y, which we get by replacing the marginal distribution in (2) with the joint distribution $p(x, y)$.

Other ways to look at MI is as a sum of entropies (3) or as the expected or average value of pointwise mutual information (4).

$$I(X;Y) = H(X) + H(Y) - H(X, Y) \tag{3}$$

$$\begin{aligned} I(X;Y) &= \mathbf{E}_{p(X,Y)} \left[i(X, Y) \right] \\ &= \sum_{x,y} p(x, y)\, i(x, y) \end{aligned} \tag{4}$$

$$i(x, y) = \ln \frac{p(x, y)}{p(x)p(y)} \tag{5}$$

Pointwise mutual information (PMI, 5) is a measure of how much the actual probability of a particular co-occurrence of events $p(x, y)$ differs from what we would expect it to be on the basis of the probabilities of the individual events and the assumption of independence $p(x)p(y)$. Note that even though PMI may be negative or positive, its expected outcome over all joint events (i.e., MI) is positive.

2.2 Mutual information in collocation extraction

Mutual information can be used to perform collocation extraction by considering the MI of the indicator variables of the two parts of the potential collocation.[3] In Table 1, I have given counts and probabilities (maximum likelihood estimates: $p = f/N$) for the collocation candidate *Mr President*, extracted from the Europarl corpus (Koehn, 2005). The MI between the two indicator variables $I(L_{\text{mr}}; R_{\text{president}})$ is in this case 0.0093. The Europarl sample consists of about 20k bigramme types with frequencies above 20. An MI of 0.0093 puts *Mr President* at rank 2 when these types are sorted after MI.

For a recent application of MI in collocation extraction see Ramish et al. (2008). More common than MI as defined above is the use of the test statistic for the log-likelihood ratio G^2, first proposed as a collocation extraction measure in

[3] In this paper, we shall use two-word collocations as our running example. The indicator variable L_w maps to *yes* when the leftmost word in a candidate is w and to *no* otherwise. Similarly for R_w and the rightmost word.

Table 1. Counts (l) and MLE probabilities (r) for the bigramme *Mr President* in a fragment of the English part of the Europarl corpus.

	$R_{\text{president}}$		
L_{mr}	yes	no	Total
yes	6 899	3 849	10 748
no	8 559	3 459 350	3 467 909
Total	15 458	3 463 199	3 478 657

	$R_{\text{president}}$		
L_{mr}	yes	no	Total
yes	.0020	.0011	.0031
no	.0025	.9944	.9969
Total	.0044	.9956	

Dunning (1993). For G^2 it has been observed that it is equivalent to MI in collocation extraction (e.g., Evert, 2005, Sect. 3.1.7).[4]

Pointwise MI is also one of the standard association measures in collocation extraction. PMI was introduced into lexicography by Church and Hanks (1990). Confusingly, in the computational linguistic literature, PMI is often referred to as simply MI, whereas in the information theoretic literature, MI refers to the averaged measure. In our example in Table 1, the bigramme *Mr President* receives a score of $i(L_{\text{mr}} = yes, R_{\text{president}} = yes) = 4.972$. In our Europarl sample of 20k types, *Mr President* comes 1573th in terms of PMI.

Although MI and PMI are theoretically related, their behaviour as association measures is not very similar. An observation often made about PMI is that low frequency events receive relatively high scores. For instance, infrequent word pairs tend to dominate the top of bigramme lists that are ranked after PMI. One way this behaviour can be understood is by looking at the PMI value of extreme cases. When two parts of a bigramme only occur together (the indicator variables of the words are perfectly correlated), we have $p(x, y) = p(x) = p(y)$. In this situation, PMI has a value of $-\ln p(x, y)$. This means that the PMI of perfectly correlated words is *higher* when the combination is *less* frequent. Even though these facts about the upper bound do not automatically mean that all low frequency events receive high scores, the upper bound of PMI is not very intuitive for an association measure.[5] Furthermore, the lack of a *fixed* upper bound means that by looking at PMI alone, we do not know how close

[4] As mentioned, we use association measures to rank candidates. A measure is thus equivalent to any monotic transformation. G^2 and MI differ by a constant factor $2N$, where N is the corpus size, if we assume a maximum likelihood estimate for probabilities (f/N), since

$$G^2 = 2 \sum_{x,y} f(x, y) \ln \frac{f(x, y)}{f_{\text{e}}(x, y)} = 2N \sum_{x,y} p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} = 2N \cdot \text{MI}$$

where the expected frequency $f_{\text{e}}(x, y) = f(x)/N \cdot f(y)/N \cdot N$.

[5] The unintuitive moving upper bound behaviour of PMI is related to the use of a ratio of probabilities. The statistical measure of effect size *relative risk* has a similar problem. Figuratively, there is a 'probability roof' that one can't go through, e.g., $p(x)$ can be twice as high as $p(y)$ when $p(y) = .05$, but not when $p(y) = .55$. The

a bigramme is to perfect correlation. In contrast, we do know how close it is to independence, since a completely uncorrelated word pair receives a PMI of 0.

A sensitivity for low frequency material is not necessarily a disadvantage. As mentioned in the introduction, different collocation extraction tasks may have different effective association measures. If we look at the MWE 2008 shared task results (Evert, 2008b), we can conclude that PMI performs relatively well as an association measure in those cases where bare occurrence frequency does not. That is, there are collocation extraction tasks in which the relative lack of a correlation with occurrence frequency is an attractive property.

MI does not suffer from a sensitivity to low frequency data, as it is an average of PMIs weighted by $p(x, y)$ – as $p(x, y)$ goes down, the impact of the increasing PMI on the average becomes less. In fact, in the kind of data we have in collocation extraction, we may expect the upper bound of MI to be positively correlated with frequency. MI equals the entropy of the two indicator variables when they are perfectly correlated. Its maximum is thus higher for more evenly distributed variables. In contingency tables from corpus data like in Table 1, by far most probability mass is in the bottom right ($L_w = no, R_v = no$). It follows that entropy, and thus maximal MI, is (slightly) higher for combinations that occur more frequently. As with PMI, however, the lack of a fixed upper bound for MI does mean that it is easier to interpret it as a measure of independence (distance to 0) than as a measure of correlation.

3 Normalizing MI and PMI

To give MI and PMI a fixed upper bound, we will normalized the measures to have a maximum value of 1 in the case of perfect (positive) association. For PMI, it is hoped that this move will also reduce some of the low frequency bias. There are several ways of normalizing MI and PMI, as in both cases the maximum value of the measures coincides with several other measures.

3.1 Normalized PMI

When two words only occur together, the chance of seeing one equals the chance of seeing the other, which equals the chance of seeing them together. PMI is then:

$$i(x, y) = -\ln p(x) = -\ln p(y) = -\ln p(x, y) \qquad (6)$$
$$\text{(when } X \text{ and } Y \text{ are perfectly correlated and } p(x, y) > 0\text{)}.$$

This gives us several natural options for normalization: normalizing by some combination of $-\ln p(x)$ and $-\ln p(y)$, or by $-\ln p(x, y)$. We choose the latter

probability roof of $p(a, b)$ is $\min(p(a), p(b)$, which, in terms of ratios, becomes further away from $p(a)p(b)$ as $p(a)$ and $p(b)$ get smaller.

option, as it has the pleasant property that it normalizes the upper as well as the lower bound. We therefore define normalized PMI as as:

$$i_{\mathrm{n}}(x,y) = \left(\ln \frac{p(x,y)}{p(x)p(y)}\right) \Big/ -\ln p(x,y). \tag{7}$$

Some orientation values of NPMI are as follows: When two words only occur together, $i_{\mathrm{n}}(x,y) = 1$; when they are distributed as expected under independence, $i_{\mathrm{n}}(x,y) = 0$ as the numerator is 0; finally, when two words occur separately but not together, we define $i_{\mathrm{n}}(x,y)$ to be -1, as it approaches this value when $p(x,y)$ approaches 0 and $p(x), p(y)$ are fixed. For comparison, these orientation values for PMI are respectively $-\ln p(x,y)$, 0 and $-\infty$.[6]

3.2 An aside: PMI^2

Since the part in the PMI definition inside of the logarithm has an upper bound of $1/p(x,y)$, one may also consider 'normalizing' this part. The result is called PMI^2, defined in (8):

$$\ln\left(\frac{p(x,y)}{p(x)(y)} \Big/ \frac{1}{p(x,y)}\right) = \ln \frac{p(x,y)^2}{p(x)(y)}. \tag{8}$$

The orientation values of PMI^2 are not so neat as NPMI's: 0, $\ln p(x,y)$, and $-\infty$ respectively. As a normalization, NPMI seems to be preferable. However, PMI^2 is part of a family of heuristic association measures defined in Daille (1994). The PMI^k family was proposed in an attempt to investigate how one could improve upon PMI by introducing one or more factors of $p(x,y)$ inside the logarithm. Interestingly, Evert (2005) has already shown PMI^2 to be a monotic transformation of the *geometric mean* association measure.[7] Here we see that there is a third way of understanding PMI^2 – as the result of normalizing the upper bound before the taking the logarithm.[8]

[6] One of the alternatives, which we would like to mention here but reserve for future investigations, is to normalize by $-\ln \max(p(x), p(y))$. This will cause the measure to take its maximum of 1 in cases of positive dependence, i.e., when one word only occurs in the context of another, but not necessarily the other way around. It seems plausible that there are collocation extraction tasks where this is a desired property, for instance in cases where the variation in one part of the collocation is much more important than in the other. See Evert (2007, Sect. 7.1), for some remarks about asymmetry in collocations.

[7] The geometric mean association measure is:

$$gmean(x,y) = \frac{f(x,y)}{\sqrt{f(x)f(y)}}$$

[8] We have further noticed that *in practice* PMI^2 is nearly a monotone transformation of X^2. To see why this may be so, consider one of the simplifications of X^2 valid in

3.3 Normalized MI

We know that in general $0 \leq I(X;Y) \leq H(X), H(Y) \leq H(X,Y)$. In addition, when X, Y correlate perfectly, it is also the case that $I(X;Y) = H(X) = H(Y) = H(X,Y)$. As in the case of PMI before, this gives us more than one way to normalize MI. In analogy to NPMI, we normalize MI by the joint entropy:

$$I_{\mathrm{n}}(X,Y) = \frac{\sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)}}{-\sum_{x,y} p(x,y) \ln p(x,y)} \tag{9}$$

MI is the expected value of PMI. Likewise, the normalizing function in NMI is the expected value of the normalizing function in NPMI: $-\sum_{x,y} p(x,y) \ln p(x,y) = \mathbf{E}_{p(X,Y)}[-\ln p(X,Y)]$.

The orientation values of NMI are 1 for perfect positive and negative correlation, and 0 for independence. It is possible to define a signed version of (N)MI by multiplying by ± 1 depending on the sign of $p(x,y) - p(x)p(y)$. This does not make a practical difference for the extraction results, however. The observed dispreferred bigrammes do typically not get very high scores and therefore do not get interspersed with preferred combinations.

3.4 Previous work on normalizing (P)MI

The practice of normalizing MI – whether as in (9) or by alternative factors – is common in data mining and information retrieval. An overview of definitions and data mining references can be found in Yao (2003). As mentioned above, PMI^2, as special case of PMI^k, was introduced and studied in Daille (1994), together with a range of other association measures. PMI^2 and PMI^3 were re-proposed as *(log frequency biased) mutual dependency* in Thanopoulos et al. (2002), in an attempt to get a more intuitive relation between PMI's upper bound and occurrence frequency.

4 A preliminary empirical investigation

To get a better feeling for the effect of normalizing MI and PMI, we will present results of evaluating NMI and NPMI against three parts of the MWE 2008 shared task.

---

the case of two indicator variables (Evert, 2005, Lemma A.2):

$$\frac{N \cdot [f(L_w = yes, R_v = yes) - f_{\mathrm{e}}(L_w = yes, R_v = yes)]^2}{f_{\mathrm{e}}(L_w = yes, R_v = yes) \cdot f_{\mathrm{e}}(L_w = no, R_v = no)}$$

It is not uncommon for $f_{\mathrm{e}}(L_w = no, R_v = no)$ to be nearly N and for $f_{\mathrm{e}}(L_w = yes, R_v = yes)$ to be orders of magnitude smaller than $f(L_w = yes, R_v = yes)$ in a co-occurrence table. If we 'round off' the formula accordingly and take its logarithm, we arrive at PMI^2.

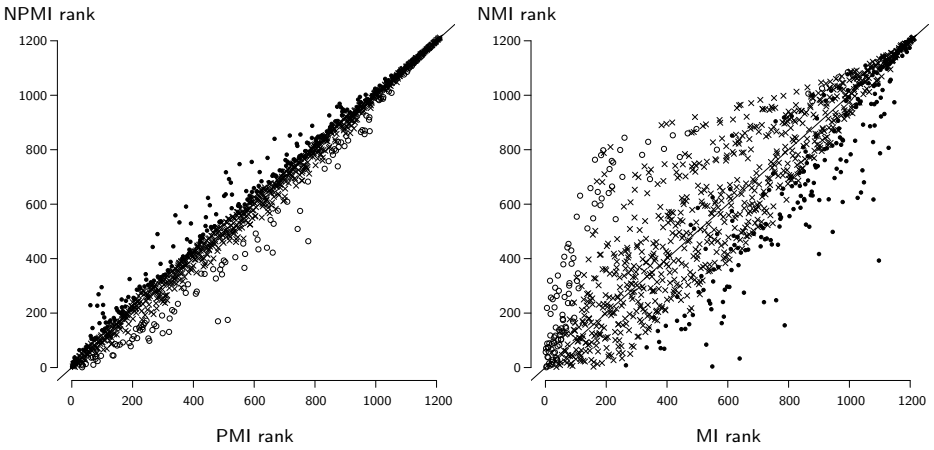

Fig. 1. Normalization of PMI (l) and MI (r) per frequency group (German AN data).

The procedure is as follows: The collocation candidates are ranked according to the association measure. These lists are then compared to the gold standards by calculating *average precision*. Average precision takes a value of 100% when all collocations are ranked before the non-collocations. Its value equals the percentage of collocations in the dataset when the candidates are randomly ordered.

The first dataset contains 1212 adjective-noun bigrammes sampled from the *Frankfurter Rundschau* (Evert, 2008a). We consider three different subtasks on the basis of this dataset, depending on how narrow we define collocation in terms of the annotation. The second dataset is described in Krenn (2008) and contains 5102 German verb-PP combinations, also taken from the *Frankfurter Rundschau*. Here, too, we look at three subtasks by considering either or both of the annotated collocation types as actual collocations. For the final and third dataset, we look at 12232 Czech bigrammes, described in Pecina (2008b).

Before evaluating (N)(P)MI against the gold standards, it is instructive to look at the effects of normalization on the ranking of bigrammes produced by each measure. To this end, we plotted the rankings according to the original measures against their normalized counterparts in Fig. 1. From the left plot, we conclude that PMI and NPMI agree well in the ranking: the ranks fall rather closely to the diagonal. For MI and NMI, to the right, we see that normalization has more impact, as the points deviate far from the diagonal.

In addition, the plotted data has been divided into three groups: high, medium, and low frequency. Normalizing PMI should reduce the impact of low frequency on ranking. Indeed, we see that the low frequency points fall above the diagonal – i.e., they are ranked lower by NPMI than by PMI, if we consider 1 to be the highest rank – and high frequency points fall below it. Normalizing MI, on the other hand, on average moves high frequency data points down and low frequency points up. All in all, we can see that in practice normalization

Table 2. Evaluation of (P)MI and their normalized counterparts on three datasets. Reported are the average precision scores in percent.

	German AN			German V-PP			Czech
Measure	cat 1	cat 1–2	cat 1–3	figur	support	both	bigrammes
random	28.6	42.0	51.6	5.4	5.8	11.1	21.2
frequency	32.2	47.0	56.3	13.6	21.9	34.1	21.8
pmi	44.6	54.7	61.3	15.5	10.5	24.4	64.9
npmi	45.4	56.1	62.7	16.0	11.8	26.8	65.6
pmi^2	45.4	56.8	63.5	17.0	13.6	29.9	65.1
mi	42.0	56.1	64.1	17.3	22.9	39.0	42.5
nmi	46.1	58.6	65.3	14.9	10.6	24.6	64.0

does what we wanted: normalizing PMI makes it slightly less biased towards low frequency collocations, normalizing MI makes it less biased towards high frequency ones.

Although not as clearly observable as the effect of normalization, the graphs in Fig. 1 also show the relation of the un-normalized measures to simple occurrence frequency. For MI, high frequency combinations tend to appear in the upper half of the ranked bigramme list. If we rank after PMI, however, the high frequency bigrammes are more evenly spread out. PMI's alleged sensitivity to low frequency is perhaps more accurately described as a lack of sensitivity to high frequency.

Table 2 contains the results of the evaluation of the measures on the three data sets. The 'random' and 'frequency' measures have been included as baselines. The reported numbers should only be taken as *indications* of effectiveness as no attempt has been made to estimate the statistical significance of the differences in the table. Also, the results do not in any sense represent state-of-the-art performance: Pecina (2008a) has shown it is possible to reach much higher levels of effectiveness on these datasets with machine learning techniques.

Table 2 shows that NPMI and PMI^2 consistently perform slightly above PMI. The trio has below-frequency performance on the German V-PP data in the 'support' and 'both' subtasks. This is to be expected, at least for PMI and NPMI. The frequency baseline is high in these data (much higher than random), suggesting that measures that show more frequency influence (and thus *not* (N)PMI) will perform better.

The behaviour of NMI is rather different from that of MI. In fact it seems that NMI behaves more like one of the pointwise measures. Most dramatically this is seen when MI is effective but the pointwise trio is not: in the German V-PP data normalizing MI has a disastrous effect on average precision. In the other cases, normalizing MI has a positive effect on average precision.

Summarizing, we can say that, throughout, normalizing PMI has a moderate but positive effect on its effectiveness in collocation extraction. We speculate

that it may be worth using NPMI instead of PMI in general. NMI, however, is a very different measure from MI, and it makes more sense to use both the original and the normalized variant alongside of each other.

5 Conclusion and future work

In this paper, we have tried to introduce into the collocation extraction research field the normalized variants of two commonly used association measures: mutual information and pointwise mutual information. The normalized variants NMI and NPMI have the advantage that their values have fixed interpretations. In addition, a pilot experimental study suggests that NPMI may serve as a more effective replacement for PMI. NMI and MI, on the other hand, differ more strongly in their relationship. As the collocation literature has shown that the effectiveness of a measure is strongly related with the task, much more and more profound empirical study is needed to be able to declare NPMI as always more effective as PMI, however.

In the experiments discussed above, we have relied on MLE in the calculation of the association scores. Since the measures are functions of probabilities, and not frequencies directly, it is straightforward to replace MLE with other ways of estimating probabilities, for instance some smoothing method. A more radical further step would be to use a different reference distribution in the association measures, i.e., to measure $p(x, y)$'s deviation from something else than $p(x)p(y)$. A change of reference distribution may, however, force us to adopt other normalization strategies.

Finally, as indicated in Section 3, there is more than one way to Rome when it comes to normalization. We hope to have demonstrated in this paper that investigating the proposed normalized measures as well as alternative ones is worth the effort in the context of collocation research.

References

Baldwin, T.: A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech (2008) 1–2

Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics **16**(1) (1990) 22–29

Cover, T., Thomas, J.: Elements of Information Theory. Wiley & Sons, New York (1991)

Daille, B.: Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques. PhD thesis, Université Paris 7 (1994)

Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics **19**(1) (1993) 61–74

Evert, S.: The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD thesis, IMS Stuttgart (2004/2005)

Evert, S.: Corpora and collocations. Extended Manuscript of Chapter 58 of A. Lüdeling and M. Kytö, 2008, *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin. (2007)

Evert, S.: A lexicographic evaluation of German adjective-noun collocations. In: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech (2008a) 3–6

Evert, S.: The MWE 2008 shared task: Ranking MWE candidates (2008b) Slides presented at MWE 2008. `http://multiword.sourceforge.net/download/SharedTask2008.pdf`.

Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit 2005. (2005)

Krenn, B.: Description of evaluation resource – German PP-verb data. In: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech (2008) 7–10

Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)

Pecina, P.: A machine learning approach to multiword expression extraction. In: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech (2008a) 54–57

Pecina, P.: Reference data for Czech collocation extraction. In: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech (2008b) 11–14

Ramisch, C., Schreiner, P., Idiart, M., Villavicencio, A.: An evaluation of methods for the extraction of multiword expressions. In: Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech (2008) 50–53

Thanopoulos, A., Fakotakis, N., Kokkinakis, G.: Comparative Evaluation of Collocation Extraction Metrics. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas (2002) 620-625

Yao, Y.: Information-theoretic measures for knowledge discovery and data mining. In Karmeshu, ed.: Entropy Measures, Maximum Entropy and Emerging Applications. Springer, Berlin (2003) 115–136