Empirical Implications on Lexical Association Measures

Brigitte Krenn
ÖFAI, Vienna, Austria
brigitte@ai.univie.ac.at

Abstract

An empirical study is presented showing how factors such as co-occurrence frequency, linguistic constraints in the candidate data and type of collocation to be identified influence the identification accuracy achieved, on the one hand, by a mere frequency-based approach and, on the other hand, by well known statistical association measures such as mutual information, Dice coefficient, relative entropy and log-likelihood statistics. The empirical results confirm the weakness of statistical measures with respect to identifying collocations from data with a high proportion of low frequency data, and reveal differences between the individual association measures depending on the class of collocations to be identified, whether they are applied to full or base form data, and whether the test samples contain low frequency data or not.

1 Introduction

In computational linguistics, a number of statistical measures have been proposed for computing the lexical association between words employing large text corpora. Some of these methods are now going to be incorporated into lexicographic workbenches in order to improve the ranking of concordances.[1] While the mathematical properties of the statistical models are well known, it is still an open question how the particular corpus employed for collocation extraction influences the identification results.

In the paper, an empirical study is presented investigating the differences in identification accuracy between statistical association measures on the one hand, and statistical association measures as well as mere co-occurrence frequency on the other hand. The following questions shall be answered:

Q1 Do the mathematical differences between the statistical association measures have significant effects when applied to German PNV-combinations?

Q2 Is there a single best statistical measure for identifying collocations?

Q3 Is there a difference between the more sophisticated statistical association measures and a simple frequency-based approach?

The particular task is identifying PP-verb collocations from German preposition-noun-verb (PNV) combinations. Preposition, noun and verb are considered to be the major lexical elements, the collocates, of the collocation. The following parameters are varied in the experiments:

1. Inflectional variation of the verbal collocate: the verb occurs either in full form or the inflectional variants are generalized to a common base form, here infinitive.
2. Occurrence frequency of the PNV-combination: three samples are distinguished, sample A – PNV-combinations with occurrence frequency equal to or larger than 10; sample B – combinations that occur 5 times or more; and sample C – combinations that occur 3 times or more; whereby the following holds: $A \subset B \subset C$.
3. Collocation type: we distinguish Funktionsverbgefüge (FVG) and figurative expressions.

Five methods for collocation identification will be examined in section 3. The statistical measures under investigation are: specific mutual information MI [Church and Hanks, 1989], the log-likelihood statistics Lgl presented in [Dunning, 1993], the $Dice$ coefficient [Smadja *et al.*, 1996], and relative entropy I as known from information theory, see for instance [Cover and Thomas, 1991]. As a control strategy, mere co-occurrence frequency $freq$ is also taken into account. MI is examined, because it was the first proposal for a statistical approach to collocations. It is still widely employed. Lgl has been introduced as a remedy for the inadequacy of MI when applied to low frequency data. $Dice$ has been introduced as an alternative to MI for identifying source language collocations and their equivalent in a target language. I has been selected by the author and is claimed to be a general alternative to Lgl.

The prerequisites for the experiments are described in section 2. It is shown how the base data are selected from the corpus (2.1). Two classes of PP-verb collocations are distinguished (2.2), and the estimates employed for collocation identification are presented (2.3).

2 Prerequisites

2.1 Construction of Base Data

To identify potential PP-verb collocations, an 8 million word portion of the Frankfurter Rundschau Corpus[2] has been automatically part-of-speech tagged and then minimal PPs have been identified employing a stochastic phrase chunker. (See [Brants, 1996] for a description of the tagger and [Skut and Brants, 1998] for a description of the chunker.) The PNV-triples are automatically[3] selected from the syntactically preprocessed corpus according to the following criteria: P and N must be constituents of the same PP, PP and V must co-occur in a sentence. The verbs in the current study are constrained to main verbs. Based on these data, two candidate sets are created: Set 1 (PNV-full-form data) consists of pairs of PNV-full-form-triples and their occurrence frequency in the extraction corpus. Set 2 (PNV-base-form data) consists of PNV-triples where the verb forms are generalized to infinitives and related occurrence frequencies.[4] In table 1, a list is presented of the ten most frequent PNV-full-form and -base-form-triples that have been found in the corpus.

PNV-full-forms	frequency	PNV-base-forms	frequency
um Uhr beginnt	379	zur Verfügung stellen	457
bis Uhr geöffnet	182	um Uhr beginnen	420
zur Verfügung stehen	174	zur Verfügung stehen	404
zur Verfügung gestellt	143	bis Uhr öffnen	196
zur Verfügung stellen	128	ums Leben kommen	195
zur Verfügung steht	115	auf Programm stehen	193
ums Leben gekommen	111	in Anspruch nehmen	192
auf Programm stehen	98	im Mittelpunkt stehen	176
in Anspruch genommen	95	auf Tagesordnung stehen	159
am Montag sagte	95	in Frage stellen	146

Table 1: The 10 most frequent PNV-full-form- and -base-form triples occurring in the extraction corpus

2.2 Selection of Test Sets

In order to test the models for collocation identification, triples which occur less than three times are deleted from the candidate sets. A cut-off threshold of three has been chosen because of the following considerations: On the one hand we want to eliminate low frequency data, because applying statistical models to low frequency data leads to poor results. On the other hand, we want to keep a certain number of low frequency data to increase the difficulty of the task for the models.

Note, reduction of the base set by word combinations which occur only once or two times reduces the set of candidate collocations by 97 % for the current corpus. The remaining 3 % (10 430 items) of PNV-triples have been manually classified into collocations and noncollocations. In addition, the collocations have been grouped into Funktionsverbgefüge (FVG) and figurative expressions employing the criteria shown in figure 1. The thus selected set of "true" collocations is used for evaluating the output of the statistical models.

For testing, the reduced set of collocation candidates is grouped into three samples A, B and C resulting in three sets with varying difficulty for the statistical models. It is expected that the models in general perform best on set A and worst on set C, as A contains only high frequency data whereas C covers a large number of low frequency data. The table below shows the total number of PNV-triples in the sets.

	A	B	C	A	B	C	
full form data	747	2 864	10 431	1 249	4 489	14 660	base form data

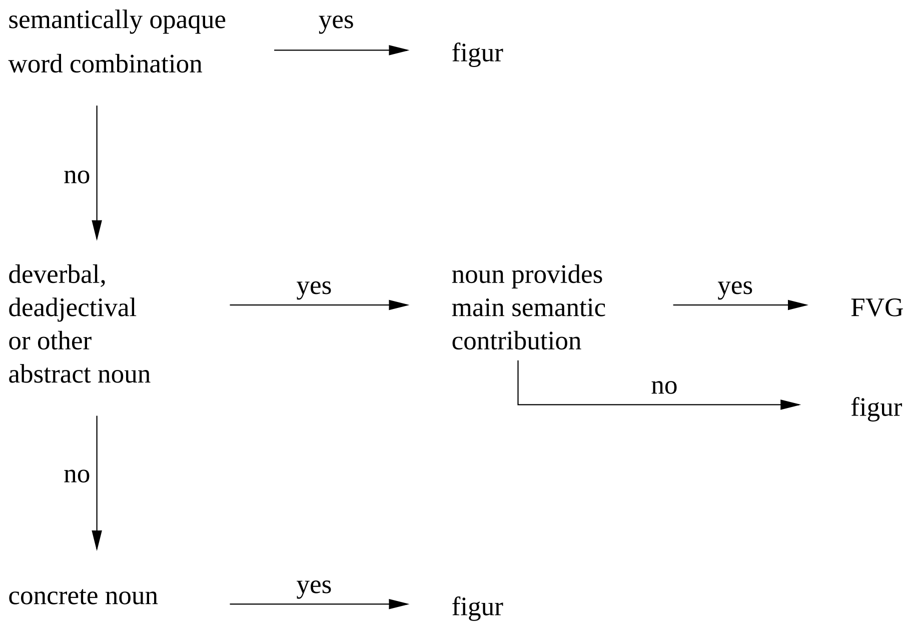

Figure 1: Criteria for manual distinction of figurative expressions and FVG

2.3 Measures and Estimates

As identification of PP-verb collocations is reduced to identifying PN-V collocations, the data required can be represented in a 2×2 contingency table. See table 2.

	c_2	$\neg c_2$
c_1	$f(c_1 c_2)$	$f(c_1 \neg c_2)$
$\neg c_1$	$f(\neg c_1 c_2)$	$f(\neg c_1 \neg c_2)$

Table 2: Contingency table for collocations with two collocates

with
$c_1 =$ PN, $c_2 =$ V,
$a = f(c1c2)$ is the frequency of a particular PNV-triple PNV$_i$, $i = 1 \ldots m$ the number of PNV-triples in the candidate set,
$b = f(c1\neg c2)$ is the frequency of all PNV-triples consisting of PN$_{PNV_i}$ but a verb other than V$_{PNV_i}$,
$c = f(\neg c1c2)$ is the frequency of all PNV-triples consisting of V$_{PNV_i}$ but a PN-tuple other than PN$_{PNV_i}$,
$d = f(\neg c1\neg c2)$ is the frequency of all PNV-triples other than PNV$_i$,
$N = a + b + c + d$,
n is the corpus size.

Employing the estimates we calculate:

$$MI = \log \frac{\frac{a}{n}}{\frac{a+b}{n} * \frac{a+c}{n}} \qquad Dice = \log \frac{2 * \frac{a}{n}}{\frac{2a+b+c}{n}}$$

$$Lgl = 2 * (\tfrac{a}{n} * \log \frac{\frac{a}{n} * N}{\frac{a+b}{n} * \frac{a+c}{n}} + \tfrac{b}{n} * \log \frac{\frac{b}{n} * N}{\frac{a+b}{n} * \frac{b+d}{n}} + \tfrac{c}{n} * \log \frac{\frac{c}{n} * N}{\frac{c+d}{n} * \frac{a+c}{n}} + \tfrac{d}{n} * \log \frac{\frac{d}{n} * N}{\frac{c+d}{n} * \frac{b+d}{n}})$$

$$I = \frac{a}{n} * \log \frac{\frac{a}{n}}{\frac{a+b}{n} * \frac{a+c}{n}} + \frac{b}{n} * \log \frac{\frac{b}{n}}{\frac{a+b}{n} * \frac{b+d}{n}} + \frac{c}{n} * \log \frac{\frac{c}{n}}{\frac{c+d}{n} * \frac{a+c}{n}} + \frac{d}{n} * \log \frac{\frac{d}{n}}{\frac{c+d}{n} * \frac{b+d}{n}}$$

Considering the above formulas, we see that MI and $Dice$ account only for positive occurrences (a, b, c)[5], whereas Lgl and I take the complete range of data into account. In addition, there is a strong similarity between Lgl and I, i.e., both measures model the relation between two frequency distributions p, q in basically the same way, namely $p \log \frac{p}{q}$. Thus we expect little difference between the latter measures in the experiments.

3 Empirical Study

3.1 Hypotheses

In order to answer questions Q_1 to Q_3, we establish two pairs of research and null hypotheses H_1, and H_0 for FVG and figurative expressions.

1. One pair for addressing the general differences between the models:

   $H_{1_{1_{FVG/figur}}}$: The lexical association models differ in their feasibility to identify FVG/figurative expressions.

   $H_{0_{1_{FVG/figur}}}$: There are no differences between the association models with respect to identifying FVG/figurative expressions.

2. One pair for examining whether there is one superior model:

   $H_{1_{2_{FVG/figur}}}$: There is a single best model for identifying FVG/figurative expressions.

   $H_{0_{2_{FVG/figur}}}$: There is no single best model for identifying FVG/figurative expressions.

3.2 Identification Procedure

For comparing the models $Dice$, I, MI, Lgl and $freq$ the following automatic procedure is employed.

1) Each model is applied to the elements in samples A, B and C. As a result each PNV-combination per sample is associated with 5 values: a *Dice*-, *I*-, *MI*-, *Lgl*- and *freq*-value.

2) For each model, the PNV-combinations per sample are ordered in descending order starting with the combination which has been assigned the highest collocativity value according to the model employed. Thus there are at most 5 different orderings for the elements of a sample.

3) From these orderings, the n highest ranking word combinations are selected, with $n = 500$, 1 000, 1 500, 2 000. Whereby n determines the borderline between collocations and noncollocations. In the case of set A, there is only one sample, $n = 500$, because set A full forms contains just 747 word combinations. The strategy has been chosen because it imposes uniform evaluation criteria on the models.

4) The thus specified samples are compared with the manually selected list of true collocations.

Summing up, first we get for each model nine samples of potential collocations: A $n = 500$ – the 500 highest ranking PNV-combinations when the model has been applied to set A, B $n = 500$ – the 500 highest ranking PNV-combinations when the model has been applied to set B, ..., C $n = 2\ 000$ – the 2 000 highest ranking PNV-combinations when the model has been applied to set C. Then the results are compared with the manually extracted list of true collocations. Thus for each sample A, B, C of size n and each model a number of correctly identified collocations and a number of false collocations is identified. These numbers are the basis for statistical significance testing.

3.3 Evaluation

First of all, we are interested in the significance of the differences between the five models *MI*, *Dice*, *Lgl*, *I* and *freq*. χ^2 tests for k independent samples are chosen as test statistics because the test is nonparametric and applicable to data at nominal scale.[7] In terms of our test data, independent samples means that each model for collocation identification selects a different subset (sample) from the candidate data. The data are at nominal scale, as PNV-combinations are grouped together according to their occurrence frequency in the extraction corpus, with occurrence frequencies being used as group labels. χ^2 tests for five independent samples are computed employing the results achieved by the (five) models for samples A $n = 500$ to C $n = 2\ 000$.

In a second step, χ^2 tests for two independent samples are computed for those samples A $n = 500$ to C $n = 2\ 000$, where a significant difference between the models has been found in the first step. Now it is tested whether there is a best model or a group of best models which significantly differ from the other model(s). "Best" is measured in terms of precision, i.e., the number of correctly identified collocations. Results are presented for FVG in table 3 and for figurative expressions in table 4.

In addition, it is distinguished whether the models are applied to full or base form data. The notation X/Y means that the models X and Y identify the same number of true collocations. X(Y) means that the difference between the number of collocations

Funktionsverbgefüge					
set	n	full forms		base forms	
		best mods	χ^2	best mods	χ^2
A	500	MI Dice 134 118	1.19 n.s.	I/Lgl freq 112 103	0.38 n.s.
B	500	freq I/Lgl 101 90	0.65 n.s.	freq I/Lgl 103 59	13.62 .001
	1 000	I/Lgl freq 201 163	4.6 .05	freq I/Lgl 159 133	2.51 n.s.
	1 500	I/Lgl MI 269 253	0.52 n.s.	I/Lgl freq 192 189	0.01 n.s.
	2 000	I(Lgl) freq 310(298) 288	0.23 n.s.	I/Lgl freq 251 210	3.92 .05
C	500	freq I(Lgl) 101 54(51)	16.16 .001	freq I/Lgl/Dice 103 4	100.51 .001
	1 000	freq I/Lgl 163 124	5.87 .02	freq Lgl 159 38	81.08 .001
	1 500	freq I/Lgl 223 205	0.79 n.s.	freq I 189 92	36.19 .001
	2 000	I/Lgl freq 298 288	0.16 n.s.	freq I/Lgl 210 134(133)	17.46 .001

Table 3: Results: the best association models for identifying FVG from PNV-full and -base forms comparing MI, $Dice$, I, Lgl, mere occurrence frequency $freq$; n.s. = not significant

identified by the models is very small. For example the information related to set C $n = 500$ in table 3 reads as follows: $freq$, I and Lgl are the best models for identifying FVG from PNV-full-forms. I and Lgl are grouped together as they identify a similar number of true collocations, i.e., 54 in the case of I and 51 in the case of Lgl. I and Lgl are opposed to $freq$ which identifies 101 true collocations. Considering the numbers 101 and 54 or 51, it is obvious that $freq$ is better than I or Lgl. This assumption is verified by employing a χ^2 test for two independent samples to the models identifying the highest number of collocations, here $freq$ and I. The resulting empirical χ^2 value is 16.16. Comparing the value with a table of theoretical values[8] shows that the difference between the models tested is highly significant ($\alpha = .001$), which we interpret as $freq$ being significantly better than I, and as a consequence significantly better than any of the other models. In the case of base form data the highest ranking model, $freq$ with 103 true collocations, is compared with the next lower ranking models which are I, Lgl and $Dice$ all three identifying just 4 true collocations. Again it is obvious from the figures that $freq$ is better than the other models. This is confirmed by $\chi^2 = 100.51$ with significance level $\alpha = .001$.

Finally, the goodness of the association models is also measured in terms of recall, i.e., the number of true collocations found by a specific model divided by the total number

Figurative Expressions					
set	n	full forms		base forms	
		best mods	χ^2	best mods	χ^2
A	500	no significant differences between the 5 models			
B	500	no signif. diff. between the 5 models		freq MI 71 62	0.56 n.s.
	1 000	no significant differences between the 5 models			
	1 500	no significant differences between the 5 models			
	2 000	I freq 208 207	0.87 n.s.	MI freq 232 202	2.17 n.s.
C	500	freq I 65 50	1.93 n.s.	freq Dice 71 41	8.46 .01
	1 000	freq I/Lgl 110 82	4.2 .05	freq Dice 121 70	14.47 .001
	1 500	freq I/Lgl 162 122	5.92 .02	freq Dice 173 95	24.29 .001
	2 000	freq I/Lgl 207 170	3.795 n.s.	freq Dice 202 112	27.38 .001

Table 4: Results: the best association models for identifying figurative expressions from PNV-full and -base forms comparing MI, $Dice$, I, Lgl and mere occurrence frequency $freq$; n.s. = not significant;

of true collocations in sets A, B or C. Table 5 presents an overview of the recall values. For each sample the model is listed which leads to the highest recall. Obviously recall increases with increasing sample size n.

3.4 Results and Interpretation

Considering table 5, we find similar patterns with respect to recall of FVG and figurative expressions from full form data versus base form data. Recall deteriorates from set A to set C in the case of full form data, whereas recall is highest from set B and second best from set A in the case of base form data. Secondly, recall of FVG and figurative expressions from base form data deteriorates more smoothly, i.e., 83 to 64 to 51 % for FVG and 69 to 55 to 38 % for figurative expressions. Recall from full form data, on the other hand, deteriorates in irregular steps, i.e., while the differences between recall from set A and B are rather small, the percentage of recall is halved from B to C.

Summing up, recall in general is worst with respect to set C. In the case of full form data, good recall can be achieved applying MI to high frequency data (set A) and I to high and medium frequency data (set B). In the case of base form data B ranges higher than A, and I or Lgl are better suited for FVG, while MI is more appropriate for figurative expressions. It is also important to notice that $freq$ cannot compete with the best statistical models when recall shall be maximized.

	full forms			base forms		
sets	A	B	C	A	B	C
FVG	144	369	710	174	304	412
best model	134	310	298	112	251	210
	(MI,	(I,	(I/Lgl,	(I/Lgl,	(I/Lgl,	(freq,
	n = 500)	n = 2 000)	n = 2 000)	n = 500)	n = 2 000)	n = 2 000)
recall	93 %	84 %	42 %	64 %	83 %	51 %
figur	96	282	586	150	338	527
best model	80	208	207	82	232	202
	(MI,	(I,	(freq,	(MI,	(MI,	(freq,
	n = 500)	n = 2 000)	n = 2 000)	n = 500)	n = 2 000)	n = 2 000)
recall	83 %	74 %	35 %	55 %	69 %	38 %

Table 5: Highest recall values of FVG and figurative expressions from set A, B and C of full and base form data

Considering tables 3 and 4, we see that there are in all cases significant differences between the models with respect to identifying FVG, but there are no significant differences between the five models when identifying figurative expressions from set A $n = 500$ and set B $n = 1\ 000, 1\ 500$ full and base forms, as well as set B $n = 500$ full forms. In other words, $\mathrm{H}_{1_{1_{FVG}}}$ is valid, whereas $\mathrm{H}_{1_{1_{figur}}}$ must be partially rejected. As a consequence, the question about the existence of a single best model or a group of best models may be asked for FVG with respect to all samples, and for figurative expression with respect to sets C, sets B $n = 2\ 000$ full forms, B $n = 500,\ 2\ 000$ base forms.

A general observation is that the performance of the statistical models is poor with respect to set C. In all cases – identifying FVG and figurative expressions from base and full form data – a mere frequency based approach either outperforms the statistical models or is at least as good as the best statistical model. I, Lgl and $freq$ however approximate with increasing sample size n in the case of full form data. The empirical results confirm what is expected according to the mathematical properties of the statistical models, namely the superiority of I and Lgl over MI and $Dice$ for samples containing low frequency data. But the results also clearly show that statistical models in general deteriorate when applied to low frequency data. The following particular tendencies could be observed,

with respect to FVG: MI and $Dice$ are the highest ranking models for identifying FVG from set A of full forms, whereas I, Lgl and $freq$ are the highest ranking models for identifying FVG from set A of base forms. I and Lgl are the best statistical models for identifying FVG from sets B and C of full and base forms. $Freq$ is always among the best models when FVG are identified from sets B and C full and base form data, and it performs best for C base forms.

with respect to figurative expressions: Other than for FVG, there is no such clearcut difference between the models in identifying figurative expressions. I, Lgl, $Dice$, MI and $freq$ are equally well suited for identifying figurative expressions from set A of full and base form data, and in the majority of cases, the models are equally well suited for identifying figurative expressions from set B of full and base form data. There is, on the one hand, a slight preference for MI in the case of base form data and, on the other hand, a slight preference for I in the case of full form data. $Freq$ clearly outperforms $Dice$, the best statistical model for set C of base form data. $Freq$ is also always among the best models for the full form data. Given set C, I and Lgl are the best statistical models.

with respect to full versus base form data: In the case of full form data compared to base form data, the numbers of true collocations identified approximate for the best models. This is due to the fact that precision of statistical association models, especially I and Lgl, strongly increases from base to full forms; but precision of $freq$ stays approximately the same.

4 Conclusion

Summing up, the results from the experiment have confirmed that there is no single best model for collocation identification, but that the performance of the models is influenced by the co-occurrence restrictions in the candidate sets, the linguistic constraints applied to the candidate data, and the class of collocations examined. Thus the following insights may be used as guidelines for collocation identification in the work of lexicographers.

With decreasing co-occurrence threshold and increasing proportion of low frequency data among the collocation candidates, precision of the statistical measures deteriorates. For an increase of identification accuracy, the following two strategies may be pursued: 1. consider only word combinations with high co-occurrence frequency, then apply statistical models, preferably MI or $Dice$ for full form data and I or Lgl for base form data; 2. on the one hand apply I or Lgl to the complete data set and select the highest ranked word combinations, on the other hand select the highest ranked word combinations according to co-occurrence frequency. Combine the two sets. While 1. leads to a stronger increase in precision, a much higher number of collocations is identified employing 2. Another advantage of 2. is that high and low frequency collocations are covered, while in the case of 1. only high frequency collocations are touched. It is also important to note that I and Lgl lead to particularly good results when the candidate set consists of full form data. In the case of base form data, $freq$ outperforms I and Lgl. As a consequence, strategy 2. is preferably applied to full form data, and strategy 1. to base form data. Full form and base form data also differ with respect to recall, i.e., high recall of FVG and figurative expressions can be achieved employing high frequency full form data or high and medium frequency full and base form data. Considering the class of collocations examined, the experimental results have revealed that MI and $Dice$ are the best models, in terms of precision, for identifying FVG from high frequency full form data, whereas I, Lgl and $freq$ are the best models for identifying FVG from high frequency base form data. On the other hand all models are equally well suited for identifying figurative expressions from high frequency base and full form data.

The results presented in the paper provide a further step towards a better understanding of empirical implications in statistical collocation identification. However, similar experiments need to be pursued employing different corpora and examining other collocations than PP-verb combinations.

References

[Brants, 1996] Thorsten Brants. TnT – A Statistical Part-of-Speech Tagger. Technical Report, Universität des Saarlandes, Computational Linguistics, 1996.

[Church and Hanks, 1989] K.W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76 – 83, Vancouver, Canada, 1989.

[Cover and Thomas, 1991] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 1991.

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

[Siegel, 1956] Sidney Siegel. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill Kogakusha Ltd., Tokyo, 1956.

[Skut and Brants, 1998] Wojciech Skut and Thorsten Brants. Chunk Tagger. Stochastic Recognition of Noun Phrases. In *ESSLI Workshop on Automated Acquisition of Syntax and Parsing*, Saarbrücken, Germany, August 1998.

[Smadja *et al.*, 1996] Frank Smadja, Kathleen R. McKeown and Vasileios Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22(1):3 – 38, 1996.

[1] Such an example is Qwick the lexicographic workbench developed by O. Mason and J. Sinclair at Birmingham University (http://www.clg.bham.ac.uk/QWICK/doc/).

[2] The Frankfurter Rundschau Corpus is part of the European Coding Initiative ECI Multilingual CD-ROM distributed by Text Encoding Initiative TEI.

[3] The candidate selection is implemented in Perl.

[4] Mmorph – the MULTEXT morphology tool provided by ISSCO/SUISSETRA, Geneva, Switzerland – has been employed for determining the infinitives.

[5] $a + b = f(c_1)$, $a + c = f(c_2)$, $a = f(c_1 c_2)$

[6] Each research hypothesis H_1 comes with a null hypothesis H_0 which is the negative formulation of the statement in H_1. Depending on the results of the test statistics either H_1 or H_0 is true.

[7] For an introduction to test statistics and details on the χ^2 test see for instance [Siegel, 1956] or any other introductory book on test statistics.

[7] Tables are available from books on test statistics.