Statistics for sentential co-occurrence

Willners, Caroline; Holtsberg, Anders

2001

Lund University, Dept. of Linguistics
Working Papers 48 (2001), 135–147

Statistics for sentential co-occurrence

Anders Holtsberg & Caroline Willners

Introduction

There is a growing trend in linguistics to use large corpora as a tool in the study of language. Through the investigation of the different contexts a word occurs in, it is possible to gain insight in the meanings associated with the word. Concordances are commonly used as a tool in lexicography, buth while the study of concordances is fruitful it is also tedious, so statistical methods are gaining grounds in corpus linguistics.

Several statistical measures have been introduced to measure the strength in association between two words, e.g. *t*-score (Barnbrook 1996:97-98), mutual information, MI (Charniak 1993; McEnery & Wilson 1996; Oakes 1998) and Berry-Rogghe's *z*-score (1973). Those measures are designed to measure the strength of association between words occurring at a close distance from each other, i.e. immediately next to each other or within a fixed window span. Research that uses the sentence as a linguistic unit of study has also been presented. For example, antonymous concepts have been shown to co-occur in the same sentence more often than chance predicts by Justeson & Katz 1991, 1992 and Fellbaum 1995.

A problem using the sentence as unit of study is that the lengths of the sentences vary from sentence to sentence. This has an impact on the statistical calculation – it is more likely to find two given words in a long sentence than in a short one. The probability of finding two given words co-occurring in the same sentence is thus affected. We introduce an exact expression for the calculation of the expected number of sentential co-occurrences. The *p*-value is calculated assuming that the number of random co-occurrences follows a Poisson distribution. A formal proof justifying this approximation is provided in the appendix.

Apart from the statistical methods that account for the variation in sentence length, a case study is presented as an application of the statistical method. The study replicates Justeson and Katz's 1991 study that shows that English antonyms co-occur sententially more frequently than chance predicts. The

results of our study show that the variation in sentence length causes the chance for co-occurrence of two given words to increase. However, the main finding of Justeson & Katz is reinforced: antonyms co-occur significantly more often in the same sentence than expected by chance.

Definitions

The terms *collocation* and *co-occurrence* are used in the literature in a somewhat inconsistent manner. Sinclair 1991:170 defines collocation as "the occurrence of two or more words within a short space of each other in a text" and co-occurrence is sometimes used synonymously. We will here let *collocation* be the occurrence of two or more words within a space of fixed length in a text, while *co-occurrence* is defined as the occurrence of two words within a linguistic unit. The linguistic unit can be a phrase, a sentence, a paragraph, an article, a corpus, etc. As indicated in the title of the paper, the statistical methods presented will focus on sentential co-occurrence, and though the case studies will concern sentential co-occurrence, the same methods could be applied to for example phrasal co-occurrence.

The tokenization of sentences is usually problematic: the period is the most common type of punctuation to end sentences, but also the most ambiguous one. For example, apart from normal punctuation it is found in numerical expressions, e.g. *13.5%;* in alphanumerical references, e.g. *5.2.4.7;* dates, e.g. *2001.01.01;* and in abbreviations, e.g. *e.g.* However, the texts we use as input have been pre-processed by a tagger, which apart from labelling the words with parts of speech, has also disambiguated the periods. A sentence is thus defined as a sequence of words ending with a punctuation tagged as a sentence delimiter. In the Brown corpus, the tag for sentence delimiters is simply '.'.

Variation in sentence length

Variation in sentence length has been extensively studied in relation to readability, cf. Björnsson 1968, Platzack 1973 and Miller 1951:124-26, 131-139 and stylistic studies (Marckworth & Bell 1967). However, there is a lack of discussion on variation in sentence length in statistical studies of co-occurrence using the sentence as unit, cf. Justeson & Katz 1991, 1992 and Fellbaum 1995.

Using sentential co-occurrence as a measure is convenient because the sentence is a well-defined unit that is usually marked in tagged corpora. But it has its drawbacks. Since the sentences in a corpus vary in length, the probability of finding two given words co-occurring in them varies as well.

The probability of finding a sentential co-occurrence or two given words must be higher in a sentence of 25 words than in a sentence of 5 words.

Think of it in terms of the urn model. Let each sentence be a ball in the urn. The sentence length is reflected by the size of the ball, i.e. small balls represent short sentences and larger balls represent longer sentences. There are 61,201 sentences in the Brown corpus, so that is the total number of balls in the urn. We also know the number of balls containing *big*, 312, and that there are 275 balls containing *little*. But we are interested in the ones containing both *big* and *little*, so we first pick out all the balls containing *big*. Among those 312 balls of different sizes there is of course a better chance to find the word *little* in one of the larger balls since there are more words in them. From this view it is even more obvious that the probability of finding two given words co-occurring in a large ball is higher than in one of the small balls.

Now, assume that all sentences are of equal length, L, i.e. all balls have the same size. If we pick out all the ones containing the first word, the probability of finding the second word among these balls is smaller than finding them in one of the balls in the urn, because the possible slots in each of the balls already picked is only $L-1$. This is also a problem we will account for.

Accounting for variation in sentence length

We assume we have a corpus with M words divided into N sentences. In the corpus there is a small number n_1 of sentences where one particular word (or lemma) occurs, as well as a small number n_2 of sentences where another word (or lemma) occurs. We observe that in a very small number x of all sentences, both words co-occur. Is this number high enough to let us conclude that co-occurrences are not only due to pure chance?

The standard way to do this is to calculate the p-value, i.e. the probability that x or more co-occurrences occurred under the null hypothesis that all co-occurrences are due to chance alone.

It has been suggested (Justeson & Katz 1991) that the number of co-occurrences of two words follows the hypergeometric distribution and that the expected number of co-occurrences is n_1n_2/N. This is, however, too crude an approximation. Here we shall derive the exact expression for the expected number of co-occurrences by taking into account both the non-uniform sentence length (which increases the co-occurrence probability substantially) and the fact that if one position in a sentence is taken by one of the words under study then that position can not also be taken by the other kind of word (which decreases the co-occurrence probability slightly).

Once we have the expected number of co-occurrences, the *p*-value is easily computed using a Poisson distribution. Many types of rare event counts follow this distribution very closely, and this approximation can be motivated theoretically in the present case since we assume that both n_1 and n_2 are small compared to N. A formal proof that this approximation is correct is given in the appendix, which involves some quite heavy mathematics. The expected number of co-occurrences is, on the other hand, rather easily calculated.

Let sentence number k have length L_k. Even though these numbers are in reality random, we shall consider them fixed, i.e. we condition the analysis on the observed sentence lengths. Statisticians call this 'conditioning on non-informative marginals'.

Let X be the number of co-occurrences, which is the number of sentences in which both words occur. At first we shall assume that no two co-occurrences are found in the same sentence so that there are n_1 words of the first kind and also n_1 sentences containing a word of the first kind, and the same for n_2. Later we shall return to this problem.

Enumerate the n_1 words of the first kind *randomly*, and also the n_2 words of the second kind. Let I_{ijk} be an indicator variable that is one if the i:th word of the first kind and the j:th word of the second kind are both found in sentence number k, and zero otherwise. The expected number of co-occurrences can be written as the sum of Nn_1n_2 terms:

$$E(X) = E\left(\sum_{ijk} I_{ijk} \right) = \sum_{ijk} E\left(I_{ijk}\right). \tag{1}$$

Note that all number of pairs (i, j) must be counted here.

There are two ways to treat the situation where one of the words occurs twice in a sentence. Either we count both occurrences, or we regard it as one co-occurrence. The first solution is thus to say that we have two co-occurrences in the same sentence. The second solution is to say that we have one co-occurrence.

In the second case we must define the sentence length as the number of words that are not one of the two kinds, plus one if there is one or more words of the first kind, plus one if there is one or more words of the other kind. The following derivation of the expected number of co-occurrences applies to both situations.

In order to compute the expected number of co-occurrences under the null hypothesis, we note that the expected value of an indicator variable is the probability that the event will occur. The term $E(I_{ijk})$ may, furthermore, be

split into the product of the probability that the word number *i* of the first kind occurs in the sentence times the probability that the word number *j* of the second kind occurs in the sentence given that the first word occurs. Under the null hypothesis of random co-occurrences we thus have that

$$(2) \qquad E(I_{ijk}) = \frac{L_k}{M}\frac{(L_k - 1)}{(M - 1)}.$$

Substituting this into (1) gives

$$(3) \qquad E(X) = \frac{n_1 n_2}{M(M-1)}\sum_{K=1}^{N} L_k(L_k - 1).$$

If all sentences are of equal length, and if we ignore the fact that there is one fewer slot for word 2 if word 1 occurs in the sentence, the expected value of the number of co-occurrences is the usual expression

$$(4) \qquad \frac{n_1 n_2}{N}.$$

Sentential co-occurrence and antonymy

It has been suggested that the reason why children learn the lexical relation between the words in an antonymous pair is that the words co-occur significantly more often than chance predicts, cf. the co-occurrence hypothesis (Charles & Miller 1989; Justeson & Katz 1991, 1992; Fellbaum 1995). Justeson & Katz 1991 have presented evidence in support of the co-occurrence hypothesis. We will here replicate their study of sentential co-occurrence of antonyms in the Brown corpus using the statistical methods presented above.

Test set and test corpus

The test set consisted of the same 35 antonym pairs that Justeson & Katz used, which had previously been identified as antonyms by Deese 1965.

As in Justeson and Katz's study, a tagged version of the Brown corpus was used as a test corpus. It is genre balanced across 15 categories and consists of 500 text extracts of about 2,000 words each.

Results

A program was written in Icon (Griswold & Griswold 1997) for calculation of the expected and actual sentential co-occurrences. As input, it takes a corpus

and a list of word pairs, and it gives as output the expected and the actual sentential co-occurrences, the probability of finding as many co-occurrences as actually found and the ratio between found and expected number of co-occurrences.

The result when using Deese's adjectives and the Brown corpus as input is given in Table 1. The individual words and their number of sentential occurrences are listed in the left part of the table. Note that it is the number of sentences that are in question, and not the total number of occurrences of a word. The right part of the table lists sentential co-occurrences. In the column *Obs.*, the observed number of sentences in which both *Adj1* and *Adj2* occur is listed. The next slot, *Exp.*, gives a figure of how many sentences with co-occurrence of the two words that is expected to be found. *Ratio* is the ratio between expected and observed co-occurrences. The last column, *Prob.*, shows the probability of finding the number of co-occurrences actually observed, or more.

Like Justeson & Katz 1991, we find that most of the antonyms exhibit sentential co-occurrence, and they are statistically significant. 25 of the 30 word pairs co-occur significantly often at the 0.05 level, 19 at the 0.01 level, and 13 using a level of 10^{-4}.

Table 2 lists expected number of co-occurrences, ratio and probability when variation of sentence length is accounted for and when it is not. The probabilities are lower when variation in sentence length is not accounted for, which has the effect that more of the co-occurrences are statistically significant. 25 word pairs co-occur significantly often at the 0.05 level, 21 at the 0.01 level and 14 at level 10^{-4}, when sentence length is not accounted for.

The expected numbers of co-occurrences are slightly higher using our measures. The ratios are consistently higher when variation of sentence length is not accounted for. However, used as a measure of the strength of the relation between the words in the antonymous pair, it must be interpreted in relation to the ratios of other word pairs. Justeson & Katz computed the overall ratio between observed and expected to 8.6. Accounting for variation in sentence length, the overall ratio is 7.0.

The results show that the variation of sentence length affects the probabilities and the expected values substantially. However, it is clear that antonym adjectives do co-occur more often than chance predicts, as the co-occurrence hypothesis suggests.

Table 1. The sentential co-occurrence of Deese's adjective pairs in the tagged Brown Corpus. Variation in sentence length is accounted for and probabilities less than 10^{-4} are rounded to 0.

Words		***Sentential occurrence***		***Sentential co-occurrence***			
Adj1	***Adj2***	***N1***	***N2***	***Obs.***	***Exp.***	***Ratio***	***Prob.***
active	passive	86	11	2	0.02	99.03	0.0002
alive	dead	57	161	2	0.20	10.21	0.0169
back	front	28	78	3	0.05	64.34	0.0
bad	good	127	694	16	1.88	8.50	0.0
big	little	312	275	13	1.83	7.10	0.0
black	white	152	250	23	0.81	28.35	0.0
bottom	top	3	70	0	0.00	-	-
clean	dirty	46	37	1	0.04	27.52	0.0357
cold	hot	137	122	8	0.36	22.42	0.0
dark	light	148	62	5	0.20	25.52	0.0
deep	shallow	84	14	0	0.03	-	-
dry	wet	54	45	2	0.05	38.55	0.0013
easy	hard	109	138	0	0.32	-	-
empty	full	63	215	1	0.29	3.46	0.2511
far	near	36	16	1	0.01	81.32	0.0122
fast	slow	32	49	1	0.03	29.87	0.0329
happy	sad	95	35	1	0.07	14.09	0.0685
hard	soft	139	59	3	0.18	17.13	0.0008
heavy	light	110	62	1	0.15	6.87	0.1355
high	low	418	138	19	1.23	15.43	0.0
inside	outside	6	38	0	0.00	-	-
large	small	351	505	26	3.78	6.87	0.0
left	right	67	214	13	0.31	42.47	0.0
long	short	522	191	12	2.13	5.64	0.0
narrow	wide	61	145	2	0.19	10.59	0.0157
new	old	1024	629	30	13.75	2.18	0.0001
old	young	629	359	17	4.82	3.53	0.0
poor	rich	101	74	7	0.16	43.87	0.0
pretty	ugly	39	20	0	0.02	-	-
right	wrong	214	113	8	0.52	15.50	0.0
rough	smooth	40	35	1	0.03	33.46	0.0294
short	tall	191	55	1	0.22	4.46	0.2009
sour	sweet	4	63	1	0.01	185.88	0.0054
strong	weak	189	29	3	0.12	25.64	0.0002
thick	thin	66	90	1	0.13	7.89	0.1191

Table 2. The sentential co-occurrence of Deese's adjective pairs in the tagged Brown Corpus. Probabilities less than 10^{-4} are rounded to 0.

Words		***Sentential co-occurrences***					
		Sentence length accounted for			***Sentence length not accounted for***		
Adj1	***Adj2***	***Exp.***	***Ratio***	***Prob.***	***Exp.***	***Ratio***	***Prob.***
active	passive	0.02	99.03	0.0002	0.02	130.72	0.0001
alive	dead	0.20	10.21	0.0169	0.15	13.43	0.01
back	front	0.05	64.34	0	0.04	84.75	0
bad	good	1.88	8.50	0	1.43	11.19	0
big	little	1.83	7.10	0	1.39	9.34	0
black	white	0.81	28.35	0	0.62	37.30	0
bottom	top	0.00	-	-	0.00	-	-
clean	dirty	0.04	27.52	0.0357	0.03	36.23	0.0272
cold	hot	0.36	22.42	0	0.27	29.50	0
dark	light	0.20	25.52	0	0.15	33.58	0
deep	shallow	0.03	-	-	0.02	-	-
dry	wet	0.05	38.55	0.0013	0.04	50.76	0.0008
easy	hard	0.32	-	-	0.25	-	-
empty	full	0.29	3.46	0.2511	0.22	4.55	0.1973
far	near	0.01	81.32	0.0122	0.01	107.53	0.0093
fast	slow	0.03	29.87	0.0329	0.03	39.37	0.0251
happy	sad	0.07	14.09	0.0685	0.05	18.55	0.0525
hard	soft	0.18	17.13	0.0008	0.13	22.56	0.0004
heavy	light	0.15	6.87	0.1355	0.11	9.04	0.1048
high	low	1.23	15.43	0	0.94	20.30	0
inside	outside	0.00	-	-	0.00	-	-
large	small	3.78	6.87	0	2.88	9.04	0
left	right	0.31	42.47	0	0.23	55.89	0
long	short	2.13	5.64	0	1.62	7.42	0
narrow	wide	0.19	10.59	0.0157	0.14	13.94	0.0094
new	old	13.75	2.18	0.0001	10.45	2.87	0
old	young	4.82	3.53	0	3.66	4.64	0
poor	rich	0.16	43.87	0	0.12	57.76	0
pretty	ugly	0.02	-	-	0.01		-
right	wrong	0.52	15.50	0	0.39	20.39	0
rough	smooth	0.03	33.46	0.0294	0.02	44.05	0.0225
short	tall	0.22	4.46	0.2009	0.17	5.87	0.1567
sour	sweet	0.01	185.88	0.0054	0.00	250.00	0.0041
strong	weak	0.12	25.64	0.0002	0.09	33.75	0.0001
thick	thin	0.13	7.89	0.1191	0.10	10.38	0.0919

Conclusion

The variation of sentence length is a problem when performing statistical measures at the sentence level. The probability of co-occurrence of two words is affected by the number of words in the sentence, a fact that has been neglected in previous studies of sentential co-occurrence. This paper presents an exact expression for the expected number of co-occurrences taking into account both the non-uniform sentence length and the fact that when a word takes up a position in a sentence, this position is filled and is not available for the other word in the co-occurring word pair. We have shown that the number of random co-occurrences can be approximated to a Poisson distribution, and calculate the *p*-value under this assumption.

The statistical methods proposed were used to replicate a study of Justeson & Katz 1991 proving that antonym adjectives co-occur significantly more often than predicted by chance. Accounting for variation in sentence length affects the expected number of co-occurrences, and the probability of finding as many co-occurrences actually observed. Justeson & Katz reported an overall ratio between observed and expected co-occurrences of 8.6, while we calculate it to 7.0. Despite the lower ratio, it is clear that antonym adjectives behave as predicted in the co-occurrence hypothesis: they do co-occur significantly more often than expected by chance.

The study above was performed on written corpora, just like the studies by Justeson & Katz 1991 and Fellbaum 1995. It is important to point out that the normal language learner is not confronted with text as input, but with spoken language. The results above do not confirm the co-occurrence hypothesis; they show that antonym adjectives tend to appear in the same sentences, but not that this facilitates the acquisition of the antonym association. To dwell deeper into this matter, a first step would be to perform the study on spoken material, preferably child-directed adult speech, to see if antonym adjectives behave similarly in spoken language. There are also other factors involved. The contexts of the co-occurring adjectives have been examined and it is clear that a word is often substituted with its antonym in repeated context significantly often (Justeson & Katz 1991).

High frequency of co-occurrence and substitution in repeated contexts may be features that help the language learner to acquire antonym association. However, we think there is more to find out from spoken corpora, like prosodic cues for example. The method presented in this paper provides a tool that gives exact statistical measures when dealing with language units that vary in length. It will be useful in further investigation of the co-occurrence of

antonyms and other types of sentential co-occurrence. There may also be applications at the word level, phrase level, paragraph level, etc., units that vary in length, like sentences.

References

Barbour, A. D., L. Holst & S. Janson. 1992. *Poisson approximation*. Oxford: Clarendon Press.

Barnbrook, G. 1996. *Language and computers. A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Berry-Rogghe, G. L. M. 1973. 'The computation of collocations and their relevance in lexical studies'. In A. J. Aitken, R. Baily & N. Hamilton-Smith (eds): *The computer and literary Studies*. Edinburgh: Edinburgh University Press.

Björnsson, C.H. 1968. *Läsbarhet*. Stockholm: Liber.

Charles, W. & G. Miller. 1989. 'Contexts of antonymous adjectives'. *Applied Psycholinguistics* 10:3, 357-375.

Charniak, E. 1993. *Statistical language learning*. Cambridge, MA: MIT Press.

Deese, J. E. 1965. *The structure of associations in language and thought*. Baltimore: Johns Hopkins Press.

Fellbaum, C. 1995. 'Cooccurrence and antonymy'. *International Journal of Lexicography* 8:4, 281-303.

Griswold, R. & M. Griswold. 1983. *The Icon programming language*. Englewood Cliffs, NJ: Prentice-Hall.

Justeson, J. & S. Katz. 1991. 'Co-occurrence of antonymous adjectives and their contexts'. *Computational Linguistics* 17:1, 1-19.

Justeson, J. & S. Katz. 1992. 'Redefining antonymy: the textual structure of a semantic relation'. *Literary and Linguistic Computing* 7, 176-184.

Marckworth, M. L. & L. M. Bell. 1967. In H. Kučera & W. N. Francis (eds): *Computational analysis of present-day American English*. Rhode Island: Brown University Press.

McEnery, A. & A. Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Miller, G. A. 1951. *Language and communication*. London: McGraw-Hill.

Oakes, M. P. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Platzack, C. 1973. *Språket och läsbarheten*. Lund: CWK Gleerup Bokförlag.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Appendix. A sketch of a proof

An explicit upper bound for the difference in total variation between the true distribution of the number of co-occurrences and the Poisson distribution will be given. The formulation is an application of the theory in chapter 2 in Barbour et al. 1992.

Suppose that the observed number of co-occurrences can be written as a sum of indicator variables, the expected value of which is

(5) $$\lambda = E(X) = \sum_{\alpha \in \Gamma} E(I_\alpha),$$

where Γ is the set of all α. Assume that we have another set of indicator variables $J_{\beta\alpha}$ that has the same distribution as I_β given $I_\alpha = 1$, that is

(6) $$L(J_{\beta\alpha}; \beta \in \Gamma) = L(I_\beta \mid I_\alpha = 1; \beta \in \Gamma).$$

A cleverly chosen set gives a probability measure that has the property that we can split the set $\Gamma_\alpha = \Gamma \setminus \alpha$ in two parts Γ_α^- and Γ_β^+ such that

(7) $J_{\beta\alpha} \le I_\beta$, if $\beta \in \Gamma_\alpha^-$

(8) $J_{\beta\alpha} \ge I_\beta$, if $\beta \in \Gamma_\alpha^+$.

Note that the inequalities say something about the outcomes – i.e. the indicator variables themselves – that is always true, which is a much stronger assumption than saying that one probability is smaller than another one. For a coupling to make sense we must have a probability measure defined simultaneously on both sets.

If we can find such a coupling, then theorem 2.C in Barbour et al. (1992) says that the distance in total variation between the true distribution and a Poisson distribution (with the same expected value λ as the true distribution) can be bounded from above by the expression

(9) $$\frac{1 - e^{-1}}{\lambda} \left(\sum_\alpha \pi_\alpha^2 + \sum_{\alpha \neq \beta} \left| Cov(I_\alpha, I_\beta) \right| \right),$$

where $\pi_\alpha = E(I_\alpha) = P(I_\alpha = 1)$.

Let the number of co-occurrences be written

(10) $$X = \sum_\omega I_{ijkl}$$

where I_{ijkl} is an indicator variable that word number i of the first kind is at word position number k in the text corpus and word number j of the second kind is at word position number l. The summation is over the set Γ, which consists of all possible combinations such that k and l belong to the same sentence. Note that the total number of terms is less than $n_1^2 n_2^2 \sum L_i^2$.

Introduce the indicator variables $J_{ijkl,i'j'k'l'}$, for given i', j', k', l', be constructed in the following way. Find the word number i' of the first kind. Swap it with the word at position k'. Find the word number j' of the second kind. Swap it with the word at position l'. Let the resulting distribution be the distribution of $J_{ijkl,i'j'k'l'}$.

It is not difficult to notice that $J_{ijkl,i'j'k'l'} = 0$ except when either $i = i'$, $j \neq j'$, $k = k'$, and $l \neq l'$ (or vice versa) or all primed quantities are different from their unprimed counterparts. For the case where $J_{ijkl,i'j'k'l'} = 0$ it is obvious that $J_{ijkl,i'j'k'l'} \leq I_{ijkl}$. In the other cases we have (after some thought) that the reverse holds, $J_{ijkl,i'j'k'l'} \leq I_{ijkl}$.

We have easily the probabilities $\pi_{ijkl}=1/M(M\text{-}1)$. It remains to compute the covariances $\mathrm{Cov}(I_{ijkl}, I_{,i'j'k'l'})$ and use the above theorem.

For the case $i = i'$, $j \neq j'$, $k = k'$, $l \neq l'$ we have that

$$(11)\quad E\left(I_{ijkl} I_{ij'kl'}\right) = \frac{1}{M(M-1)(M-2)}$$

$$(12)\quad C\left(I_{ijkl}, I_{ij'kl'}\right) = \frac{1}{M(M-1)(M-2)} - \left(\frac{1}{M(M-1)}\right)^2 < \frac{1}{(M-3)^3}$$

$$(13)\quad \text{No. of terms} = n_1 n_2 (n_2 - 1) \sum_1^N L_i (L_i - 1)(L_i - 2) < n_1 n_2^2 \sum_1^N L_i^3$$

and similar for $i \neq i'$, $j = j'$, $k \neq k'$, and $l = l'$.

If all primed quantities are different from their unprimed counterparts then

$$(14)\quad E\left(I_{ijkl} I_{i'j'k'l'}\right) = \frac{1}{M(M-1)(M-2)(M-3)}$$

$$(15)\quad C\left(I_{ijkl}, I_{ij'kl'}\right) = \frac{1}{M(M-1)(M-2)(M-2)} - \left(\frac{1}{M(M-1)}\right)^2 < \frac{6}{(M-3)^5}$$

$$(16)\quad \text{No. of terms} < n_1^2 n_2^2 \left(\sum L_i^2\right)^2 .$$

For all other combinations we have

$$(17)\quad E\left(I_{ijkl} I_{i'j'k'l'}\right) = 0$$

(18) $$\left|C\left(I_{ijkl}, I_{i'j'k'l'}\right)\right| = \left| -\left(\frac{1}{M(M-1)}\right)^2 \right| < \frac{1}{(M-3)^4}$$

(19) No. of terms $< n_1 n_2 (n_1 + n_2)\left(\sum_1^N L_i^2\right)^2 + n_1^2 n_2^2 \sum_1^N L_i^3$.

Finally,

(20) $$\sum_\Gamma \pi_{ijkl}^2 = \sum_\Gamma \left(\frac{1}{M(M-1)}\right)^2 \leq \frac{n_1^2 n_2^2}{(M-3)^4} \sum L_i^2$$

(21) $$\sum_{\alpha \neq \beta} \left|Cov\left(I_\alpha, I_\beta\right)\right| <$$

$$n_1 n_2 \left(\frac{n_1 + n_2}{(M-3)^3} \sum_1^N L_i^3 + \frac{6 n_1 n_2}{(M-3)^5} \left(\sum_1^N L_i^2\right)^2 + \frac{n_1 + n_2}{(M-3)^4} \left(\sum_1^N L_i^2\right)^2 + \frac{n_1 n_2}{(M-3)^4} \sum_1^N L_i^3 \right) .$$

Now let the number of words in the corpus be large and let $n_1 n_2 / M$ be bounded as well as $\frac{1}{M}\sum L_i^3$, $\frac{1}{N}\sum L_i^2$ and the mean sentence length $M/N = \frac{1}{N}\sum L_i$. If furthermore n_1 and n_2 are of the order $\sqrt{M}$ we see that the rate of convergence to the Poisson distribution is no slower than

(22) $$d_{TV} = O\left(\frac{1}{\sqrt{M}}\right) .$$

If no more than one co-occurrence is counted in every sentence then the behaviour in the limit of the distribution is the same. This is a simple consequence of the fact that the probability of more than one co-occurrence tends to zero faster than the probability of one co-occurrence in the above model.