Table of Contents

Fix Me! **This page was translated from German into English using DeepL.**
(remove this paragraph once the translation is finished)

Significance measures in the assessment of co-occurrences

In statistics, significance is a key figure that describes the probability of a systematic correlation between variables, i.e., in the case of text analyses, between subtexts (e.g., words). The significance expresses whether an apparent connection could be purely coincidental nature or with high probability actually exists.

Depending on the object of investigation, different formulas are used for the calculation, which originate primarily from computational linguistics. The significance measures should help to separate important from unimportant cooccurrences. Statistical parameters, such as corpus size, frequency of individual words or frequency of co-occurrence, are put into relation.

One of the simplest significance measures is a frequency-sorted co-occurrence list, i.e., the frequency of co-occurrence of two words in the entire corpus. A disadvantage of frequency-sorted lists is that according to Zipf's law, the beginning of quantitative linguistics, very many words occur very rarely. Consequently, a threshold greater than 1, i.e., multiple co-occurrences of a word pair, can be used to filter out about two-thirds of the co-occurrences. Calculated by the eAQUA tools, this looks as follows for selected corpora:

corpus	number of co-occurrences	co-occurrences freq = 1	in percent
BTL [1)]	137.486.214	110,876,836	80,65
MPL [2)]	580.247.568	398.935.822	68,75
Perseus Shakespeare [3)]	6.746.602	5.027.170	74,51
TLG [4)]	355.021.014	258.961.566	72,94

As can be seen from the small overview, a large part of the cooccurrences found can rather be described as low-frequency. In order to filter out the important ones, calculation methods are required, some of which are presented here.

Dice

The Dice coefficient (also Sørensen-Dice coefficient, named after the botanists Thorvald Sørensen and Lee Raymond Dice) indicates the similarity of two terms by means of a number between 0 and 1. Basis of calculation are so-called N-grams. With N-grams, a term or a text is divided into equally sized fragments. These fragments can be letters, phonemes, whole words or similar. The number of N-grams present in both terms is determined in order to set them in relation to the total number of N-grams. It is calculated according to the formula $dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$ where n_{ab} is the intersection of both terms and n_a or n_b is the number of N-grams formed per term.

Example 1: **Term a = Tür** **Term b = Tor**	$dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$
Bigram	Trigram
a = { §T, Tü, ür, r§ } b = { §T, To, or, r§ } $d_{T\square r, Tor} = \frac{2 \times 2}{4+4} = \frac{4}{8} = 0,5$	a = { §§T, §Tü, Tür, ür§, r§§ } b = { §§T, §To, Tor, or§, r§§ } $d_{T\square r, Tor} = \frac{2 \times 2}{5+5} = \frac{4}{10} = 0,4$

Example 2 **Term a = Spiegel** **Term b = Spargel**	$dice_{ab} = \frac{2 \times n_{ab}}{n_a + n_b}$
Bigram	Trigram
a = { §S, Sp, pi, ie, eg, ge, el, l§ } b = { §S, Sp, pa, ar, rg, ge, el, l§ } $d_{Spiegel, Spargel} = \frac{2 \times 5}{8+8} = \frac{10}{16} = 0,625$	a = { §§S, §Sp, Spi, pie, ieg, ege, gel, el§, l§§ } b = { §§S, §Sp, Spa, par, arg, rge, gel, el§, l§§ } $d_{Spiegel, Spargel} = \frac{2 \times 5}{9+9} = \frac{10}{18} \square 0,556$

When evaluating co-occurrences, the dice coefficient can be used by relating the frequencies (frequencies) of the words. Here, n_a and n_b are the frequencies of the terms, n_{ab} is the number of common occurrences. The above calculation results in relatively simple evaluation scales. The more frequently the two terms are used together, the more the value approaches 1. If the two terms only occur together, the highest significance is achieved with 1. How often this co-occurrence is found in the corpus is irrelevant. This results in an important property of the Dice coefficient: Cooccurrences that rarely occur together, where one word is high-frequency and the other is low-frequency, are scored as nonsignificant.

Jaccard

The Jaccard coefficient (after the botanist Paul Jaccard) indicates the similarity of two terms by means of a number between 0 and 1. The calculation basis for text mining methods are so-called N-grams. With N-grams, a term or a text is broken down into equal parts. These fragments can be letters, phonemes, whole words or similar. The number of N-grams present in both terms is determined in order to set them in relation to the total number of N-grams. It is calculated according to the formula $jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$ where n_{ab} is the intersection of both terms and n_a or n_b is the number of N-grams formed per term.

Example 1: **Term a = Tür** **Term b = Tor**	$jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$
Bigram	Trigram
a = { §T, Tü, ür, r§ } b = { §T, To, or, r§ } $d_{T\square r, Tor} = \frac{2}{4+4-2} = \frac{2}{6} \square 0,334$	a = { §§T, §Tü, Tür, ür§, r§§ } b = { §§T, §To, Tor, or§, r§§ } $d_{T\square r, Tor} = \frac{2}{5+5-2} = \frac{2}{8} = 0,25$

Example 2 **Term a = Spiegel** **Term b = Spargel**	$jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$
Bigram	Trigram

Example 2 **Term a = Spiegel** **Term b = Spargel**	$jaccard_{ab} = \frac{n_{ab}}{n_a + n_b - n_{ab}}$
a = { §S, Sp, pi, ie, eg, ge, el, l§ } b = { §S, Sp, pa, ar, rg, ge, el, l§ } $d_{Spiegel,Spargel} = \frac{5}{8+8-5} = \frac{5}{11} \approx 0,455$	a = { §§S, §Sp, Spi, pie, ieg, ege, gel, el§, l§§ } b = { §§S, §Sp, Spa, par, arg, rge, gel, el§, l§§ } $d_{Spiegel,Spargel} = \frac{5}{9+9-5} = \frac{5}{13} \approx 0,385$

For the evaluation of co-occurrences, the Jaccard coefficient is similar to the Dice coefficient. Both calculate the significance value similarly, the relative order of the co-occurrences remains the same, only the absolute significance value differs marginally. A model calculation with mean frequency of 100 looks as follows:

n_a	n_b	n_{ab}	**Dice**	**Jaccard**
100	100	1	0,01	0,005
100	100	10	0,1	0,05
100	100	50	0,5	0,33
100	100	90	0,9	0,82
100	100	100	1	1

Poisson measure

An approach to computing significant co-occurrences is based on the Poisson distribution (named after the mathematician Siméon Denis Poisson), a discrete probability distribution $p(n,k) = \frac{1}{k!}\gamma^k e^{-\gamma}$

On the basis of the Poisson distribution give *Quasthoff / Wolff* [5] the Poisson measure with the formula. $p\left(n_a, n_b, k, n\right) = \frac{k \times (\log k - \log \gamma - 1)}{\log n}$ which has been used, for example, to compute corpora in the Vocabulary Portal, and in which the two factors **n** (number of sentences in the corpus) and **k** (frequency of co-occurrence, also called n_{ab}) are relevant.

After a conversion and the basic assumption $\gamma = \frac{n_a \times n_b}{n}$ we get the following calculation

$$p = \frac{n_{ab} \times \log \frac{n_{ab} \times n}{n_a \times n_b} - n_{ab}}{\log n}$$

Thus, the Poisson measure could be reduced to the difference between Local Mutual Information and Frequency.

Log-likelihood measure

One of the most popular significance measures in the analysis of large text corpora is according to *Dunning* [6] the log-likelihood measure, which is based on the binomial distribution, one of the most $p(K=k) = p^k (1-p)^{n-k} \binom{n}{k}$ important discrete probability distributions.

Dunning finally arrives at the formula when calculating **log likelihood**:

$$-2\log\lambda = 2\left[\log L\left(p_1,k_1,n_1\right)+\log L\left(p_2,k_2,n_2\right)-\log L\left(p_1,k_1,n_1\right)-\log L\left(p_2,k_2,n_2\right)\right]$$

under the condition

$$\log L(p,n,k) = k\log p+(n-k)\log(1-p)$$

The log-likelihood measure can thus be derived as follows

$$lgl = 2\left[n\log n-n_a\log n_a-n_b\log n_b+n_{ab}\log n_{ab}+\left(n-n_a-n_b+n_{ab}\right)\log\left(n-n_a-n_b+n_{ab}\right)+\left(n_a-n_{ab}\right)\log\left(n_a-n_{ab}\right)+\left(n_b-n_{ab}\right)\log\left(n_b-n_{ab}\right)-\left(n-n_a\right)\log\left(n-n_a\right)-\left(n-n_b\right)\log\left(n-n_b\right)\right]$$

Characteristic of the log-likelihood measure, in contrast to the Poisson measure for example, is the equal treatment of significantly frequent and significantly rare events. Thus, in the digitized data from TLG in version TLG-E, there are about 1.3 million co-occurrences in about 73.8 million words that occur only once and yet are assigned an lgl value of 30 and a little more. A similarly large value of 34.553 has, for example, **καὶ** and **τὸ**, which together were counted 14311 times.

1) Bibliotheca Teubneriana Latina, Online-Version, Status as of February 2014.

2) Patrologia Latina Database, CD-ROM Version, November 1995.

3) William Shakespeare in Perseus Digital Library, Renaissance Materials, Status as of May 2013.

4) TLG-E, CD-ROM Version from 1999.

5) [Quasthoff 02]. Uwe QUASTHOFF, Christian WOLFF. The Poisson Collocation Measure and its Applications. In Second International Workshop on Computational Approaches to Collocations, 2002.

6) [Dunning 93]. Dunning, T. “Accurate Methods for the Statistics of Surprise and Coincidence.” In: Computational Linguistics 19, 1 (1993), 61-74.