The Mantel Test

The following sections are taken from: Kevin Stadler & Matt Spike: *Measures of compositionality for artificial language experiments* (in preparation).

1 Introduction

The 'measure of structure' made popular by Kirby et al. (2008, p.10686) is very much in the spirit of Montague's definition of compositionality as a homomorphism between structured signal and meaning spaces. Crucially, this measure bypasses the question of how individual (simple) meanings map onto signals. Instead, it quantifies the regularity of the mapping between the internal structures of the signal and meaning spaces.

Assuming an experimental design in which a participant produces one signal for each of n different (complex) meanings, one can construct two distance matrices of size $n \times n$. These matrices capture the pairwise distances between all of the attested signals and meanings according to some predefined distance metrics over strings and (complex) meanings respectively.

Valid distance metrics will give rise to matrices which are symmetric, with 0s along their diagonal. Based on these distances, one can then calculate the correlation between the $n \cdot (n-1)/2$ pairs of meaning and string distances on one side of the matrix' diagonal.

As a first step, this correlation coefficient captures the degree to which *pairs of meanings which are similar to each other* map onto *pairs of signals which are similar to each other.* While the theoretical range of the coefficient is from -1 to 1, there is an additional problem in establishing whether an empirically determined level of correlation is in fact statistically significant. The fact that the data points underlying its calculation are not independent (the $n \cdot (n-1)/2$ pairs of values are derived from just n signal-meaning pairs) means that there is potential for autocorrelation in the data and that the theoretical null distribution of the underlying correlation coefficient can not be used. To account for this, Kirby et al. (2008) fall back onto a correlation significance test based on randomising locations to determine spatial autocorrelation of measures in ecology (see Cornish, 2011, p.91 for an overview of its use in linguistics). The basic idea is that, by repeatedly randomising the mapping from signals to meanings (equivalent to shuffling the columns and rows of one of the distance matrices in the same way) and calculating the respective correlations for these randomised distance matrices, one can generate an empirical null distribution of the correlation coefficient *for that particular distance matrix.* Under the assumption that the randomised correlations are normally distributed, one can then determine a z score for the actual correlation coefficient, which is nothing but the correlation coefficient's position relative to the randomised distribution expressed in standard deviations from its mean. This in turn can be used to get an estimate of the significance level of the established correlation, by determining the corresponding p value based on the Normal distribution.

This randomisation and computation of the z score are really a second step to determine the *significance* of the correlation coefficient obtained earlier. Nevertheless it is this z score, rather than the raw correlation, which has established itself as a measure of compositionality in analyses of artificial language learning tasks (Kirby et al., 2008; Carr et al., 2016; Beckner et al., 2017). It should be stressed that the z score is related to the *significance level* of a measure, rather than expressing the *effect size* of the measure itself. High z scores therefore do not necessarily capture high *levels* of compositionality, rather than high (statistical) confidence in the presence of *some* level of compositionality (Spike, 2016, p.186).

2 Analysis of the ILM data from Beckner et al. (2017)

To illustrate some of the properties of the Mantel test and its measures, we perform an in-depth analysis of the experimental data kindly provided by Beckner et al. (2017). Beckner et al.'s data is a replication of the classic iterated learning experiment setup of Kirby et al. (2008).

2.1 Choice of correlation coefficient

One degree of freedom in performing the Mantel test that is not often discussed is the choice of correlation coefficient to use. Since Brighton et al. (2005), Pearson's product-moment correlation coefficient has been the de-facto standard in the artificial language experiment literature even though its assumptions about the underlying data, in particular normally distributed values without ties, are not met by either the string nor the meaning distance data. For the meaning distance in particular it should be noted that with the experiment's fixed $3 \times 3 \times 3$ meaning space, and meaning distance being calculated as the number of dimensions in which two meaning combinations differ, the $27 \cdot 26/2 = 351$ pairwise meaning distances from which the correlation coefficient is computed can only take up one of three values, and their frequency is fully determined by the experimental design: 81 meaning distances of 1, 162 of distance 2 and 108 of distance 3.

An example of the raw distance data to which the correlation coefficients are fit can be seen in Figure 1. With the large amount of ties exhibited by the data one would expect that an ill-suited correlation measure might have a detrimental effect on the reliability of the results. But while the absolute size of Pearson's r differs from those of other correlation coefficients accounting for ties, the choice of coefficient does not appear to significantly affect the normalised z scores computed from them (see the graphs in the appendix for an exhaustive comparison of different correlation coefficients).

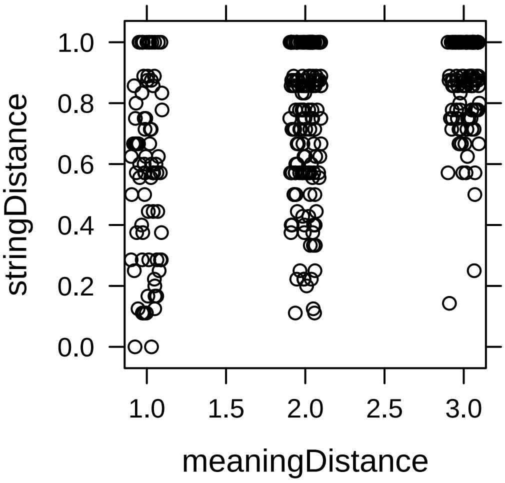


Figure 1: Example of the raw data that goes into the calculation of the correlation coefficient. Data shown is from the last generation of chain 1 of Beckner et al. (2017)'s *large* (training size 15) condition. Data is jittered along the x axis. The different correlation coefficients for this data set are Pearson's $r = 0.39$, Spearman's $\rho = 0.36$, Kendall's $L = 0.29$.

2.2 Comparison of raw correlation coefficient against normalised z score

To illustrate the relationship between the raw signal/meaning distance correlations and the derived z scores, Figure 2 plots the two against each other for the total 24 chains (across two conditions) reported by Beckner et al. (2017) with computations based on the signal-meaning pairs available for each generation. What is striking is that the two measures align perfectly in the majority of cases. It should be noted that no effort has been made to scale the two y axes to produce this overlap, including the near-perfect alignment of the 0 marks. The alignment is simply a result of independently scaling each of the y axes so that the range of plotted values covers the entire available plotting space. What this tells us is that when we are looking to compare signal-meaning space correlations between data sets that are identically sized data sets (as is the case in controlled experiments with a constant number of productions), the raw correlation measures are actually just as comparable between experiments as normalised z scores.

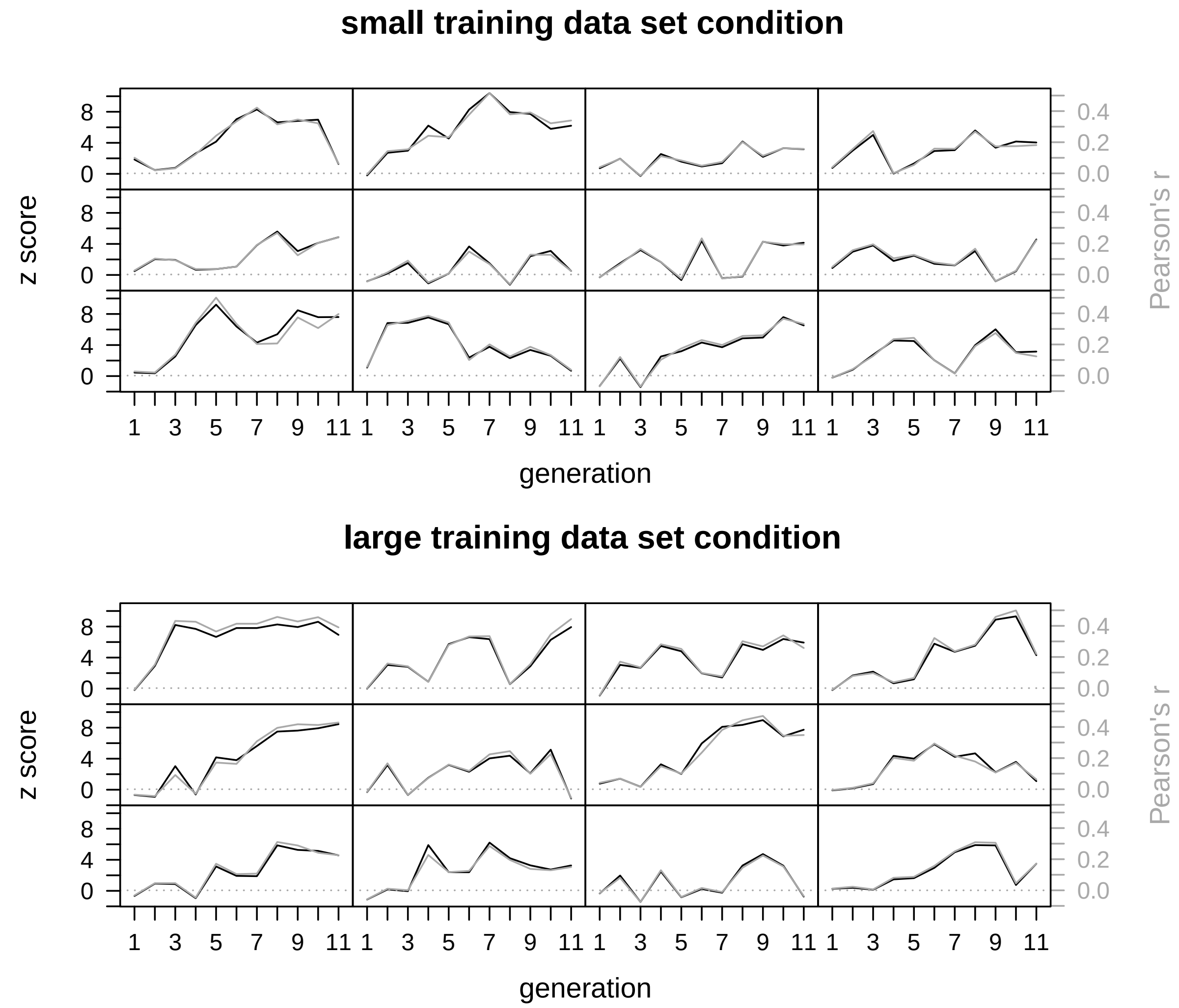


Figure 2: Comparison of the z score as derived from the distribution of rs for 1000 random permutations (black line, left axis) plotted against the raw value of Pearson's r (gray line, right axis) for the 24 iterated learning chains run by Beckner et al. (2017). The dotted gray line indicates the baseline of the correlation coefficient ($r = 0$) where there is no evidence for correlation (either positive or negative).

While the z score transformation does not appear to affect the relative levels of compositionality much, it makes the absolute level of the measure both difficult to interpret as well as difficult to compare between experiments. As mentioned previously, the z score really captures the significance level of the correlation, rather than the actual structure preserved by the signals' mapping between the form and meaning spaces. As a consequence, differences in the *size of the test sets* obtained by different experimental designs or conditions – or even by testing differences between generations – can therefore have an adverse effect on the interpretability of the measure (Cornish et al., 2009). To illustrate this point, Figure 3 shows the raw correlation as well as z score for two data sets: the first is simply the data from one of the chains already plotted above (chain no. 2 in the 'small' condition), while the measures in the central panel are based on a data set where, for every generation, each of the 27 signal-meaning pairs was duplicated. Data sets of this form occur in a variation of the iterated learning model where, instead of having just one learner per generation, there are two participants first learning the language system and then interacting. The output of one such generation of learners might therefore be not one but two tokens produced for every meaning combination. As a result, the correlation coefficient of such an experiment with doubled data size would be computed based on distance matrices of size 54 × 54.

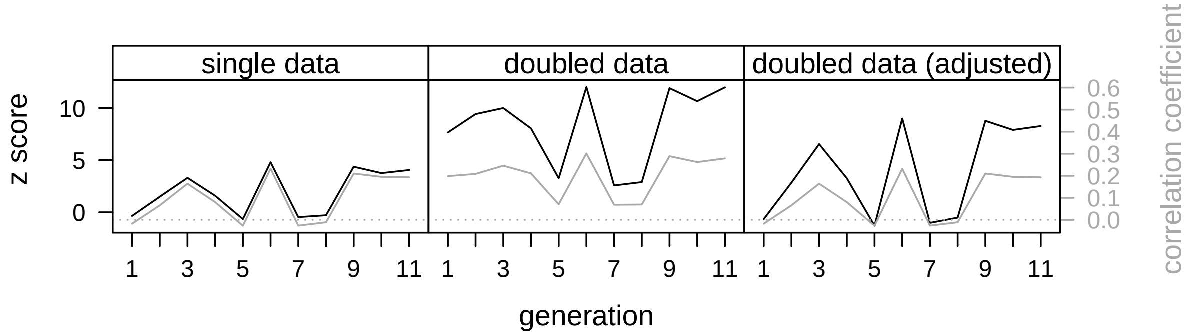


Figure 3: Comparison of the raw correlation and z score for the data set from iterated learning chain no. 7 of the small learning data condition. Left: scores calculated from the original data set (all pairwise distances between 27 signal-meaning pairs, $N = 351$). Middle: scores calculated from a duplicated data set containing all signals twice (all pairwise distances between 54 signal-meaning pairs, $N = 1431$). Right: scores calculated from the duplicated data set under omission of pairs of signals from the distance matrix which have identical meaning components ($N = 1404$).

Despite the fact that both data sets specify the exact same mappings between forms and meanings (only with twice the amount of data in the latter case), the test results differ in two aspects:

1. Both the raw r and z score exhibit a constant shift upwards from the 0 baseline.
2. Moreover, the z score sees a linear boost (of about a factor of two) relative to the raw correlation coefficient.

The latter effect is easily explained by what was just discussed: the randomised sample of correlation coefficients obtained from shuffling a larger data set of the same limited number of data points has a *lower* standard deviation, which leads the normalised z scores to exhibit a relative increase. This effect is expected (and desired) for a measure of significance, but not indicative of an actual increase in compositionality. This suggests that the z score should be categorically avoided for drawing comparisons between compositionality levels of different-sized data sets.

While the present analysis might suggest that the relationship between the raw correlation and the z score is just be one of linear scaling, we can easily determine that this is not true: any correlation measure reaches its maximum value at 1, whereas the theoretical maximum value of the z score increases as a function of its sample size (alongside other factors, in particular the pool of forms and meanings that the distance computations are based on). This same expansion of the z score scale is also problematic at the lower end of the scale, where z scores of greater than 1.96 or 1.645 (corresponding to the .05 significance level for two-sided and one-sided significance tests respectively) have been adopted as a benchmark of compositionality. But particularly in combination with the fact that the distance metric approach underlying the Mantel Test does not allow one to pinpoint which individual segments map onto which individual meanings, and the vast pool of character sequences which are candidates for such mappings, Spike (2016) has highlighted the risk of type I errors:

"[T]he Mantel score which is typically used to measure the compositionality of model and experimental data is potentially severely compromised when systems exhibit duality of patterning. As shown above, inevitable random correlations between form and meaning spaces result in highly significant, but reasonably small correlations even when systems are completely holistic." (Spike, 2016, p.195).

The other notable difference between the first two panels of Figure 3 is that not just the z score, but also the raw correlation measure increases, with its attested minimum value for the given data set jumping from 0 to almost 0.2. This is a consequence of the naive approach to increasing the amount of data by simply doubling all signal-meaning pairs: the pairwise comparison between the duplicated data points introduces pairs of 0 meaning and 0 string edit distances into the distance matrices, which inflates the consequent computation of the correlation coefficient. Crucially, the same issue arises in the analysis of experimental designs in which more than one data point is sampled for a given meaning combination, such as in the case of interacting participants. When the overall degree of compositionality is computed based on the productions of all participants pooled, identical productions by different participants would inflate the compositionality score when really they are an indication of the fact that the participants have successfully *aligned* their communication systems.

In order to disentangle the effects of convergence in multiple-participant designs from those of compositionality, some researchers have adjusted the computation of the correlation coefficient to exclude cells of the pairwise distance matrices which have a meaning distance of zero. The resulting scores using this approach can be seen in the right-most panel of Figure 3, showing that the raw correlation coefficients are indeed identical to the ones found in the original data set, while the z scores are still relatively inflated.

3 Summary

The Mantel Test is a significance test for a minimum amount of compositionality (Spike, 2016) – beyond that, using z is unnecessary and confusing, just report r – its values and boundaries are much more well-understood.

What to do with repeated measures for the same meaning combination is still an open question. Identical repeated data points inflate the measure, but does that mean that they should simply be excluded?

A general problem of the test is that the distance measure for the string and meaning spaces needs to be defined in advance. While string edit distances are well defined, the individual contribution of the different meaning dimensions is not necessarily predictable, sometimes making it necessary to establish them post-hoc.

To some extent the same is true for the string distances also: the Mantel Test doesn't actually take the content of individual signals into account, and therefore has no means of expressing what the actual 'segments' that are supposed to make up the compositional system are (Tamariz and Smith, 2008). This approach is therefore also affected by uneven morpheme lengths or shared phonotactic material. Measured levels are also affected by orthogonal features of the communication systems such as the size of the character inventory etc.

References

Beckner, C., Pierrehumbert, J. B., and Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*.

Brighton, H., Smith, K., and Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, 2(3):177–226.

Carr, J. W., Smith, K., Cornish, H., and Kirby, S. (2016). The Cultural Evolution of Structured Languages in an Open-Ended, Continuous World. *Cognitive Science*.

Cornish, H. (2011). *Language adapts: exploring the cultural dynamics of iterated learning*. PhD thesis, The University of Edinburgh.

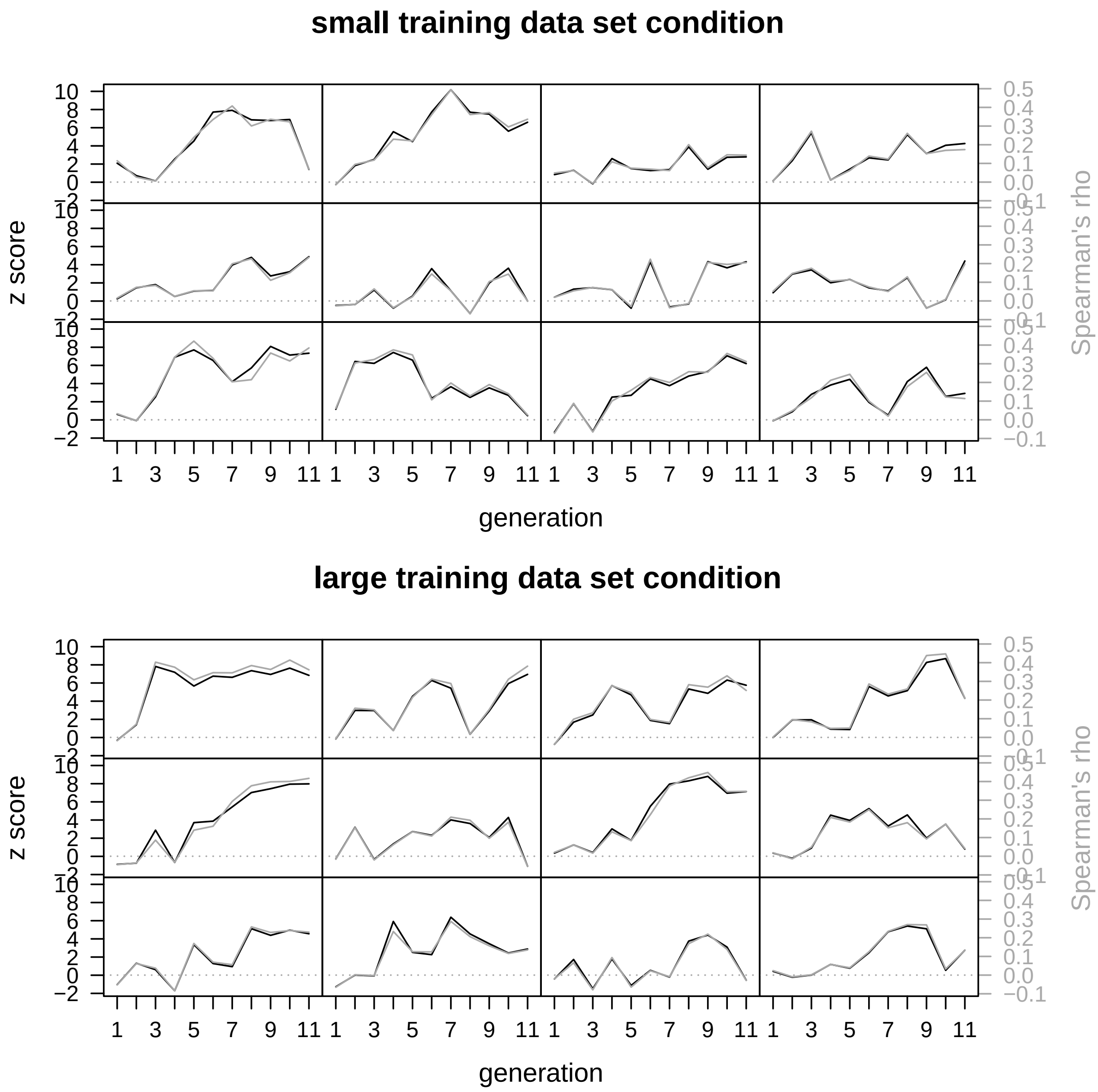
Cornish, H., Tamariz, M., and Kirby, S. (2009). Complex adaptive systems and the origins of adaptive structure: what experiments can tell us. *Language Learning*, 59:Suppl.1(December):187–205.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31):10681–10686.

Spike, M. (2016). *Minimal requirements for the cultural evolution of language*. PhD thesis, The University of Edinburgh.

Tamariz, M. and Smith, A. D. M. (2008). Regularity in mappings between signals and meanings. In Smith, A. D. M., Smith, K., and Ferrer i Cancho, R., editors, *The Evolution of Language: Proceedings of the 7th International Conference (EVOLANG7)*, Singapore. World Scientific Press.

Appendix: Comparison of results based on Pearson's, Spearman's and Kendall's correlation coefficients



small training data set condition

large training data set condition

Figure 4: Relationship between raw correlation coefficient and z score, using Spearman's ρ correlation coefficient.

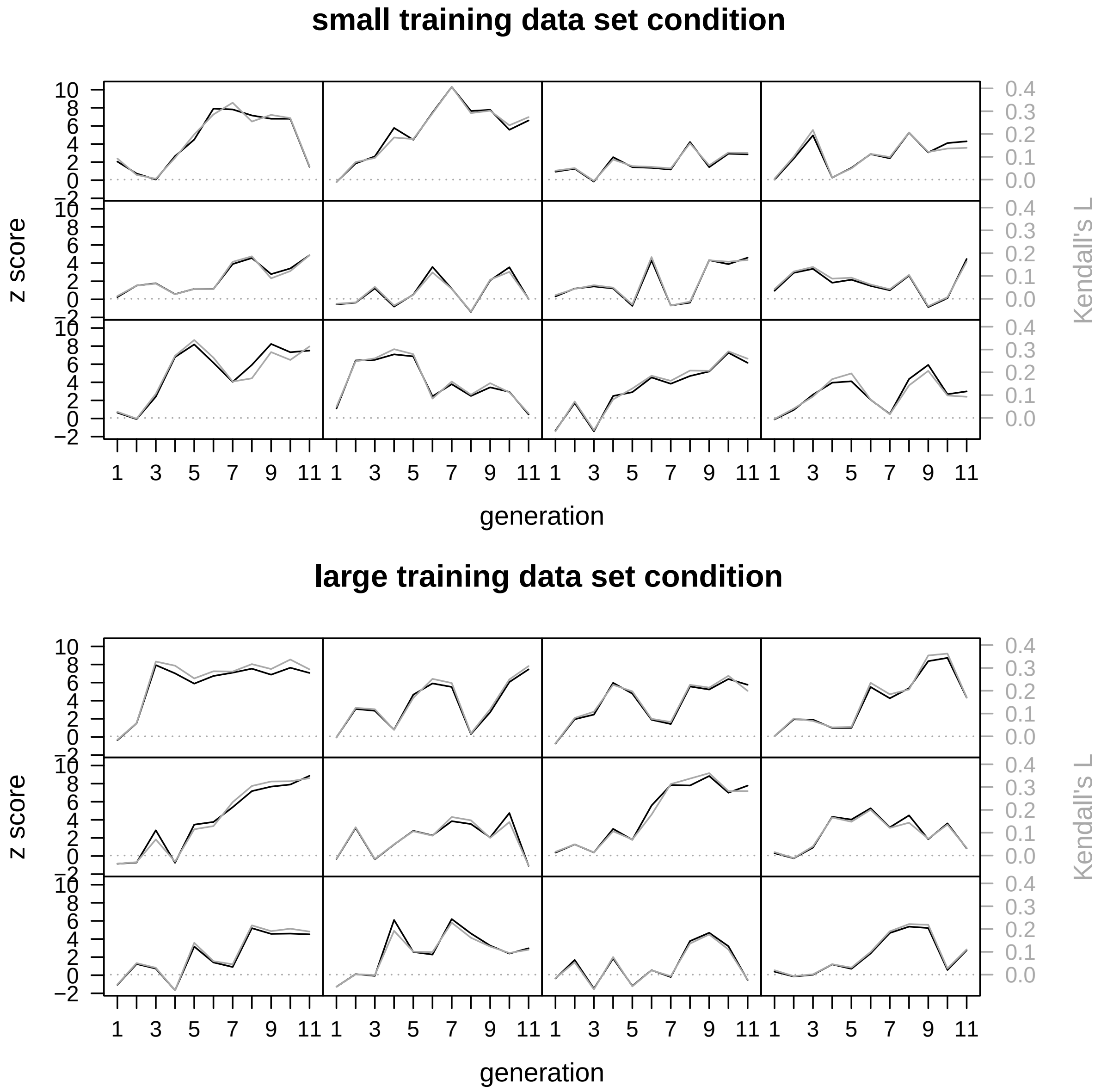


Figure 5: Relationship between raw correlation coefficient and z score, using Kendall's L correlation coefficient.

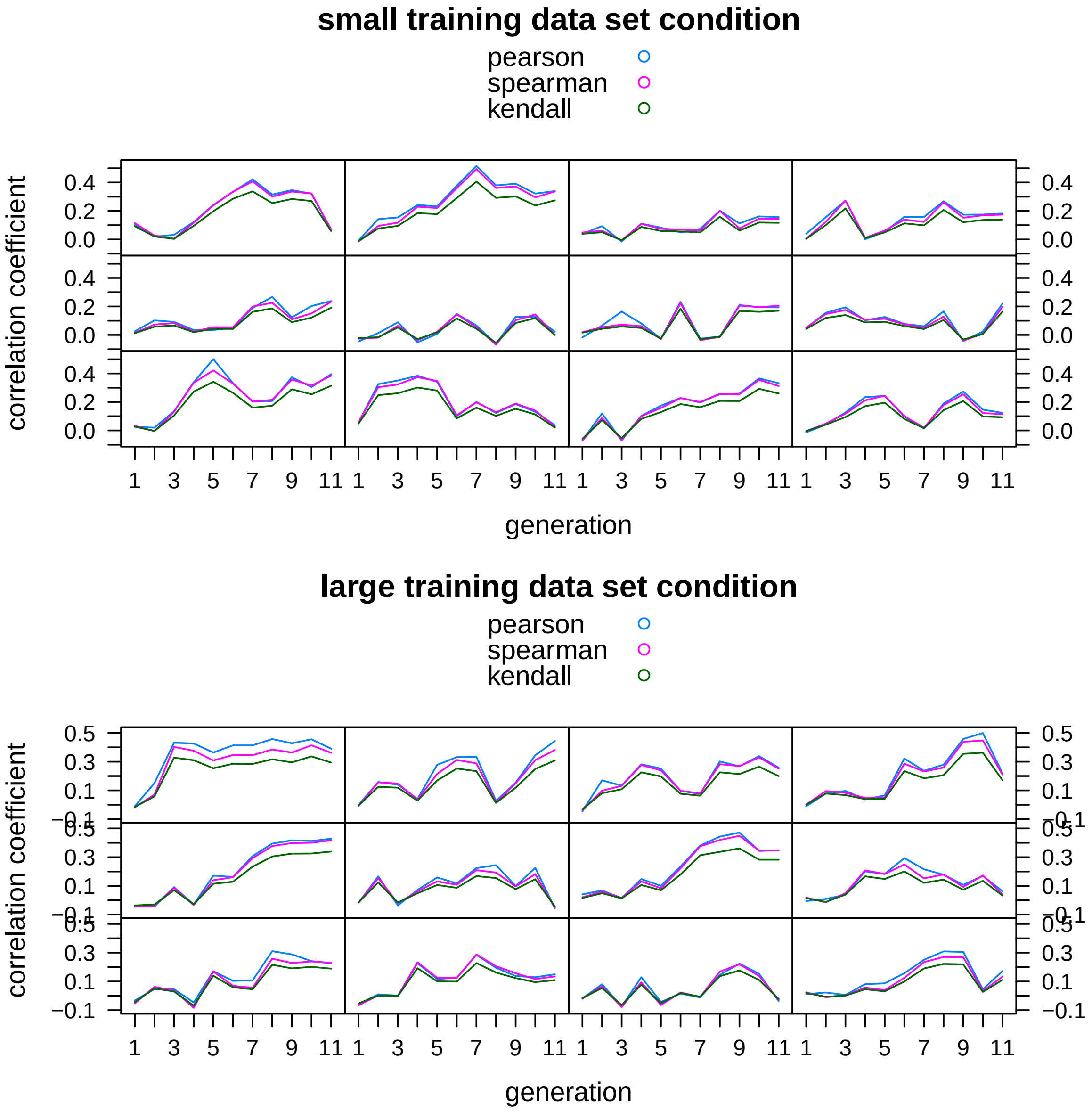


Figure 6: Comparison of the different correlation measures.

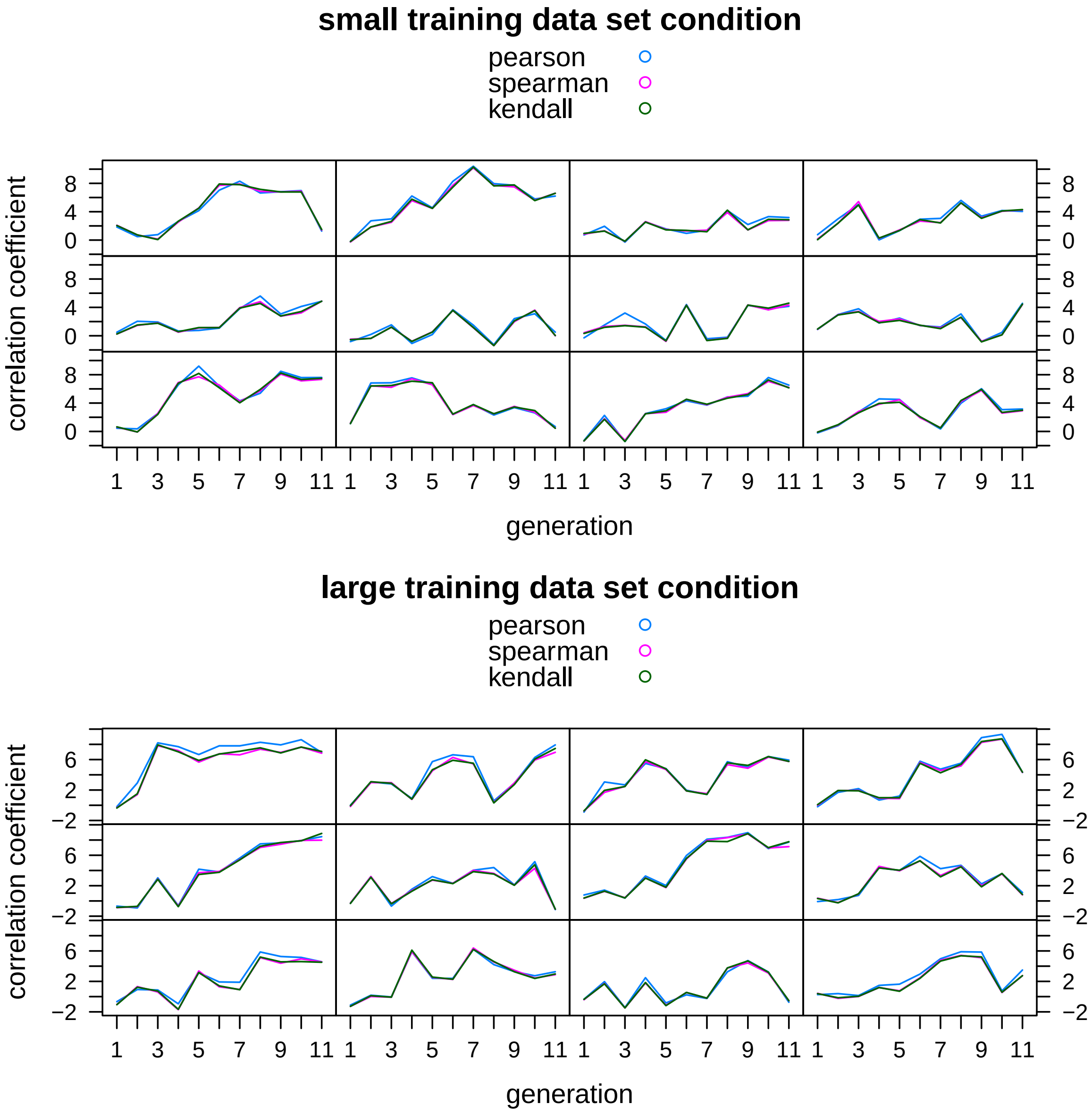


Figure 7: Comparison of z scores computed from the different correlation measures.