The Dimensionality of Language

Isidoros Doxas (doxas@colorado.edu)
Center for Integrated Plasma Studies
University of Colorado, Boulder, CO, 80309, USA

Simon Dennis (Simon.Dennis@adelaide.edu.au)
School of Psychology
University of Adelaide, SA 5005, Australia

William Oliver (oliver@colorado.edu)
Institute of Cognitive Science
University of Colorado, Boulder, CO, 80309, USA

Abstract

The dimensionality of the paragraph space of five corpora of different languages (English, French, modern Greek, Homeric Greek and German), genres (fiction and non-fiction) and intended audiences (children, adolescents and adults) is investigated. Term by paragraph occurrence data is processed by whitening, and the correlation dimension is calculated. All five corpora exhibit a weave-like structure, where at short distances the correlation dimension is lower than at long distances. In each case, the lower range has dimensionality of approximately eight. The higher range varies from about twelve to about twenty eight. Control simulations in which word instances were permuted do not exhibit two separate dimensionalities, demonstrating that the effect is determined by specific word choice, rather than by the paragraph length or word frequency properties of the corpora. By the embedding theorem (Takens, 1981), these results imply that at the lower range the trajectory can be describe by between nine and seventeen ordinary differential equations, placing an important constraint on the way in which authors transition from idea to idea when constructing prose, which may be universal.

Keywords: language, dimensionality, latent semantic analysis, correlation dimension.

In a typical vector-space model of language, any span of text can be modeled as a vector derived from the frequency with which terms occur in the texts. Various methods, like Latent Semantic Analysis (LSA, Landauer, Foltz, & Laham, 1998), can be used to reduce the dimensionality of the initial vector space, but even then the dimensionality of the space is usually quite large (of the order of 300 in LSA applications (Landauer, Foltz, & Laham, 1998), for example. However, the points that represent the paragraphs in this 300-dimensional space do not fill the space; rather they lie on a subspace with a dimension lower than the embedding dimension. The fact that a low-dimensional structure can be embedded in a higher dimensional space is routinely used in the study of nonlinear dynamical systems, where the embedding theorem (Takens, 1981) relates the dimensionality of the dataset under study to the dimensionality of the dynamics that describes it.

There are a number of different dimensions that can be defined for a given dataset, e.g. the capacity dimension, the information dimension, the Hausdorff dimension, etc. (Lichtenberg & Lieberman, 1992). A usual choice for small datasets is the correlation dimension (Grassberger & Procaccia, 1983) because it is more efficient and less noisy when only a small number of points is available. It can be shown that $D_{capacity} \geq D_{information} \geq D_{correlation}$, but in practice almost all attractors have values of the various dimensions that are close to each other (Lichtenberg & Lieberman, 1992; Grassberger & Procaccia, 1983).

The correlation dimension is derived by considering the correlation function

$$C(l) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1 (j \neq i)}^{N} H(l - |\vec{X}_i - \vec{X}_j|) \qquad (1)$$

where $\vec{X}_i$ is a vector pointing to the location of the ith point in the data set, N is the total number of data points, and H is the Heaviside function. The correlation function is therefore the number of distances that are less than l. The correlation dimension, ν, is then given by the relation

$$\lim_{l \to 0} C(l) \propto l^{\,\nu} \qquad (2)$$

The correlation dimension, as well as all other dimensions, are strictly defined only at the limit $l \to 0$. In practice, the limit means a length scale that is much smaller than any other length scale of the system. With that definition in mind, one can envision geometric structures that exhibit different well-defined dimensions at different length scales, as long as those length scales are well separated. To calculate the correlation dimension one plots the log of the number of distances less than l against the log of l. For portions of the resulting curve that are well described by a straight line, the gradient of this line is the correlation dimension. To obtain a geometric appreciation of the correlation dimension imagine that points are aligned evenly on a straight line (see Figure 1). Now consider an expanding hypersphere around one of the points. As the radius of the sphere increases the number of points within will increase linearly. Alternatively, if the points are arranged evenly in two dimensions then as the radius increases the number of points will increase quadratically, and so on.

A typical example is a long pipe, which appears one- two- or three-dimensional depending on the lengthscale that we are considering (cf. Figure 1). At small scales points fill a three dimensional space. At intermediate scales (defined by the thickness of the tube wall and the diameter of the tube) points fill a two dimensional space and at large scales the pipe is one dimensional. A similar example in the reverse order is a large piece of woven cloth, which looks two dimensional at long scales, but is composed of one-dimensional threads at short scales. This is the picture presented by the corpora we have studied; they look low-dimensional at short scales and higher-dimensional at long scales.

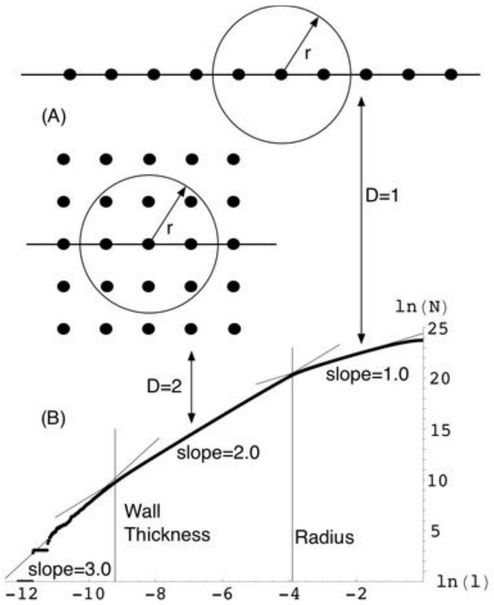

Figure 1: The measured dimensionality of a long pipe. (A) A schematic representation of the scaling of the correlation function with distance; the number of points withind a distance r scales as r^D. (B) The correlation function for 100,000 points randomly distributed in a tube of length unity. The radius of the tube is 10^{-2} and the thickness of the tube wall 10^{-4}. The slopes give dimensions of 3.0, 2.0, and 1.0 respectively at length scales that are smaller than the thickness of the wall, between the thickness of the wall and the diameter of the tube, and longer than the diameter of the tube.

We calculated the correlation dimension of five corpora, in English, French, modern and Homeric Greek, and German (see Appendix for more details of how the spaces were prepared). The English corpus includes text written for children as well as adults, representing the range of texts that a typical US college freshman will have encountered. The French corpus includes excerpts from articles in the newspaper Le Monde, as well as excerpts from novels written for adults. The modern Greek corpus is comprised of articles in the political, cultural, economic, and sports pages of the newspaper Eleftherotypia. The German corpus includes articles from German textbooks and text extracted from internet sites, and is intended to represent the general knowledge of an adult native speaker of German. The Homeric corpus consists of the complete Iliad and Odyssey. The documents in all five corpora are paragraphs (stanzas for Homer), most of which are 80-500 words long. The English corpus includes 37651 paragraphs, the French 36126, the German 51027, the modern Greek 4032, and the Homeric 2241 paragraphs (stanzas).

Fig. 2A shows the log of the number of distances, N, that are less than l plotted against the log of l for the English corpus. The slopes give dimensions of 8.3 and 19.5 for the short and long distances respectively. Figures 2B, 2C, 2D, and 2E show the same plot for the French, Greek, German and Homeric corpora respectively. The slopes give a short-distance dimension of 8.8, 8.4, 7.4, and 8.6, and a high dimension of 12.4, 28.0, 22.3, and 26.6 respectively. All five corpora clearly show a "weave-like" structure, in which the dimensionality at short distances is smaller than the dimensionality at long distances.

As a control, we also calculated the correlation dimension for a space constructed by randomly combining words from the English space. Fig. 3 shows a plot of the correlation function for that corpus. The number of documents, the length of each document, and the numbers of occurrences of each word are the same in the random and original corpora, but the random corpus does not have a low-dimensional structure. Instead the points are space filling within the limitations of the sample size. This implies that the observed low-dimensional structure is a property of the word choice in the paragraphs, and not a property of the word frequency or paragraph length distributions of the corpus.

As indicated in the appendix, we retained between 300 and 420 dimensions when calculating the plots above, which is typical of the number of dimensions that show optimal performance in applications of Latent Semantic Analysis to essay grading. Figure 4 shows the impact of varying the number of dimensions retained for the English corpus. As the number of components increases, the two scale structure becomes more pronounced and the slopes converge.

The main reason we are usually interested in knowing the dimensionality of a dataset is the embedding theorem (Takens, 1981), which relates the dimensionality of a dataset to the dynamics that generated it. The theorem states that (almost) all datasets of dimension d can be described by at most $2[d] + 1$ Ordinary Differential Equations (ODE's), where $[d]$ is the integer value of d, but in practice the situation is more favourable. Most well known systems are actually described by the minimum $[d] + 1$ equations, like the Lorentz, Rössler, or Ueda systems, which have dimensions slightly over two, and are described by three equations (Sprott, 2003).

In light of this, it is worth noting that the measured short-scale dimensionalities of the five corpora are not only surprisingly low but also suggestively similar. This allows one to hope that the short-scale language dynamics can be described by nine ODE's, while the embedding theorem virtually guarantees that it can be described by at most seventeen. The fact that five such diverse corpora exhibit the same dimensionality at short distances is encouraging, since it implies that at short semantic distances all languages might be described by the same nine equations.

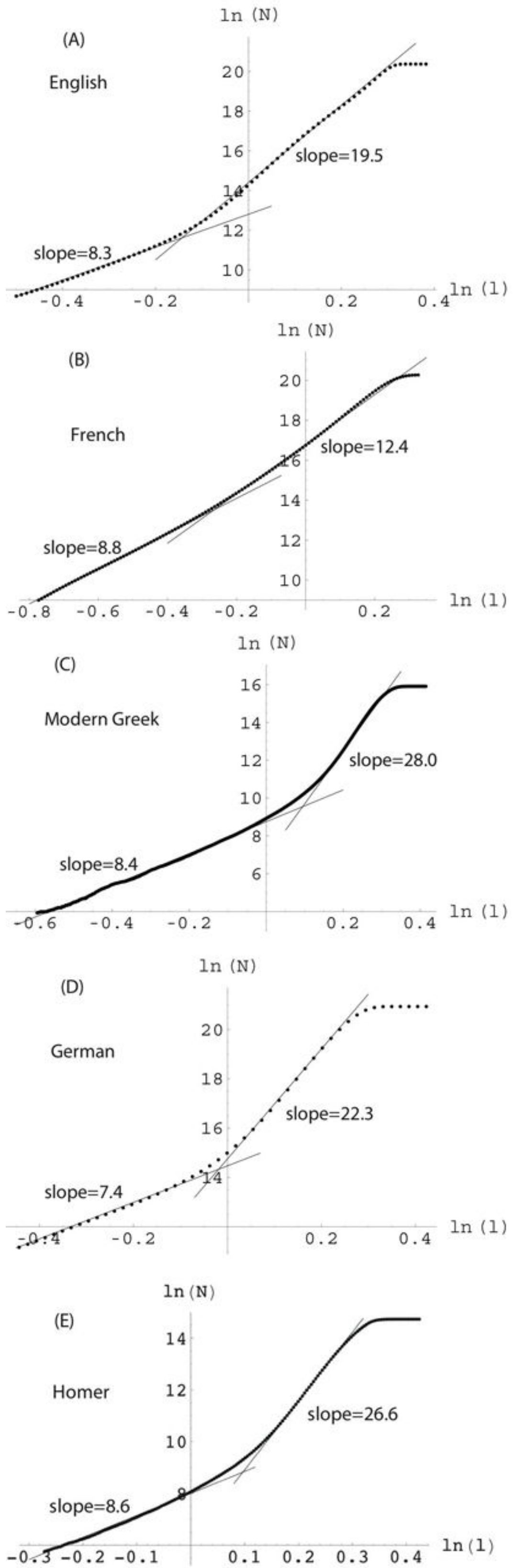

Figure 2: The measured dimensionality of the five corpora. (A) the English corpus, (B) the French corpus, (C) the modern Greek corpus, (D) the German corpus, (E) Homer. All corpora exhibit a low-dimensional structure, with the dimensionality at long scales being higher than at short scales.

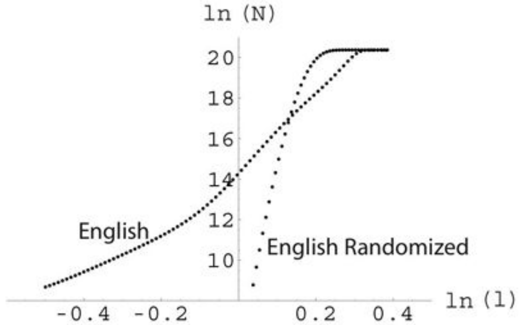

Figure 3: The measured dimensionality of the randomized English corpus. The randomized corpus does not show the low dimensional structure of the English corpus, and it is space filling within the limitations of the number of points used. This implies that the low-dimensional structure is a property of the word choice in the paragraphs, and not of paragraph length or word frequency in the corpus.

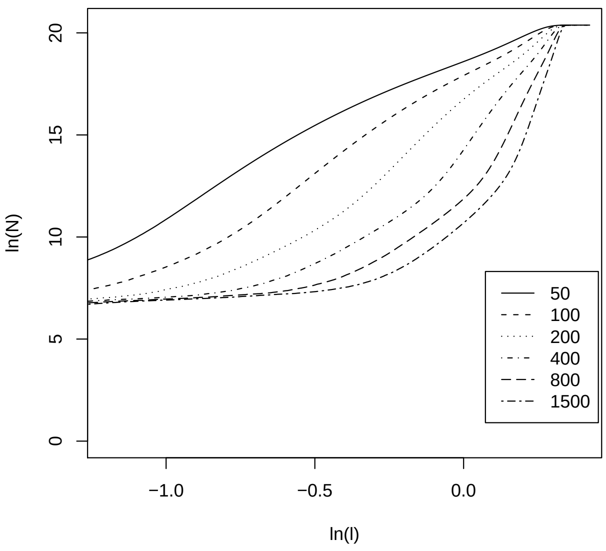

Figure 4: As the number of dimensions retained increases the two scale structure becomes more pronounced and the slopes converge. The plot above shows this progression for the English corpus.

Implications for Models of Language Trajectories

In addition to pointing at a universal low-dimensional dynamics of language, the above results can also guide the development of models of language. Perhaps the simplest way one could attempt to characterize the paragraph trajectory would be as a Gaussian random walk in semantic space, and such a model is implicitly assumed in applications of Latent Semantic Analysis to the testing of textual coherence (Foltz, Kintsch, & Landauer, 1998) and to textual assessment of patients for mental illnesses such as schizophrenia (Elvevag, Foltz, Weinberger, & Goldberg, 2001). However, a Gaussian random walk model of that type cannot describe an attractor with two distinct dimensionalities at different length scales; the trajectory such a model describes will fill the space in which it takes place.

So what generates the two scale structure? An examination of document pairs in the two distance ranges suggests that the short distances correspond to paragraph pairs that treat the same subject and come from the same larger context (e.g. neighbouring paragraphs from the same article or book chapter) while longer distances correspond to paragraph pairs that treat clearly different subjects. This observation suggests a hierarchical model (Blei, Griffiths, Jordan & Tenenbaum, 2004) in which a discourse can be modeled by a Gaussian random walk within a given topic, while topic transitions follow different statistics. This process can be modeled by a Levy-like walk in which the rare long jumps take place in more dimensions than the numerous short ones.

In order to simulate the basic characteristics of a hierarchical topics model of this kind, we constructed a 3-dimensional geometric structure that consists of randomly oriented 2-dimensional square planes. Figure 4 shows plots of this structure for two different values of the parameter R_l, which is the ratio of the size of the planes to the scale of the system. Figures 4A and 4C show plots of the correlation function of these structures for $R_l = 1/2$ and $R_l = 1/8$ respectively, while Figures 4B and 4D show a sample of 4000 points for the two structures respectively. We see that the dimensionality plots which resemble the plots produced by the corpora correspond to R_l values that describe a dense but granular geometric structure, in the sense that the 2-dimensional planes are clearly distinct but are not isolated. When the 2-dimensional planes are small enough to be isolated (i.e. the ratio R_l is small) the dimensionality plots show a plateau at the length scales that are not well populated. The range of R_l values over which the models produce the observed behaviour is relatively small, suggesting that the observed dimensionality can place important constraints on the model.

Conclusions

The above results were obtained with corpora which, in addition to being in different languages, were constructed for different intended audiences, represent widely different genres (including 3000 year old poetry), and are of different lengths. Despite this diversity, the

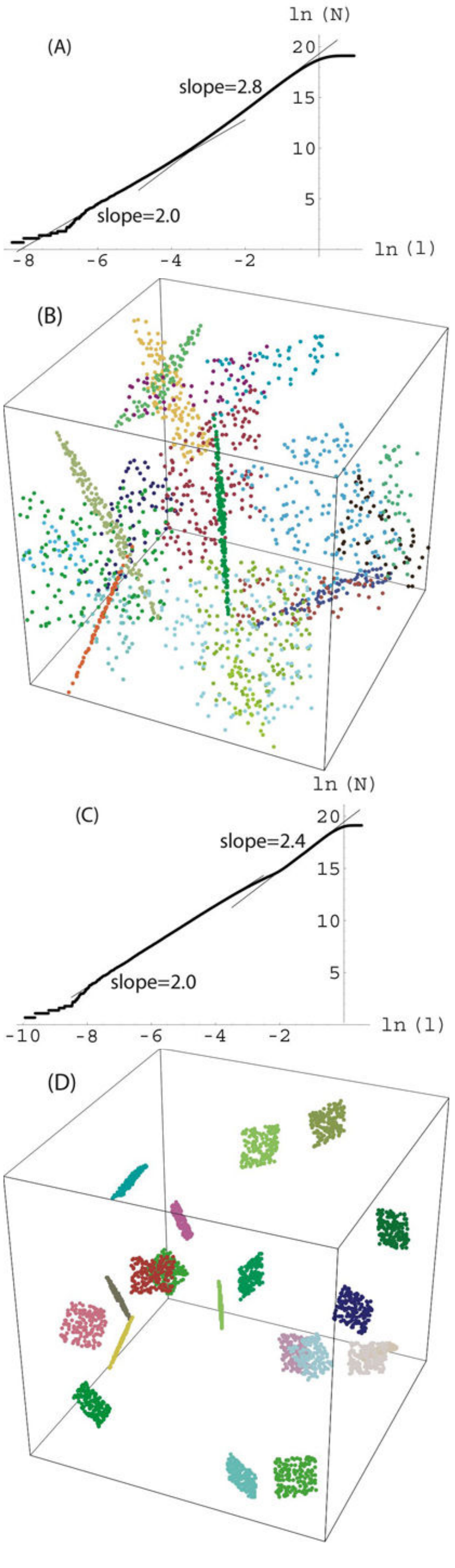

Figure 5: The measured dimensionality of a Levy flight hierarchical topics model. (A) and (C) show the correlation function, (B) and (D) show a sample of 4000 points. (A) and (B) were obtained with $R_l = 1/2$, (C) and (D) with $R_l = 1/8$. For the lower ratio the dimensionality plot shows a clear plateau at the intermediate scales that are not well populated. The plateau becomes increasingly more pronounced as the ratio decreases.

results show a surprising common simplicity in the structure of human language, which is encouraging for the development of future quantitative models, and argues for a universal language dynamics.

Acknowledgments

We thank the late Stephen Ivens and Touchstone Applied Science Associates (TASA) of Brewster, New York for making available the English corpus. We also thank Wolfgang Lenhard of the University of Wüzburg for permission to use the German corpus. We are indebted to the Perseus Digital Library of Tufts University for making the complete Homer available on the web. This work was partly supported by grants from the National Science Foundation and the US Department of Education.

References

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, **25**, 295-284 (1998).

Takens, F. (1981). Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence*, David Rand and Lai-Sang Young, editors. Lecture Notes in Mathematics, **898**, 366. Springer, Berlin.

Lichtenberg, A. J., & Lieberman, M. A. (1992). *Regular and Chaotic Dynamics* (Second Edition). Applied Mathematical Sciences **38**, Sprigner-Verlag, New York.

Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D* **9**, 189.

Sprott, J. C. (2003). *Chaos and Time-Series Analysis*, Oxford University Press, Oxford, pp. 329-336.

Foltz, P., Kintsch, W., and Landauer, T. K., (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes* **25**, 285-307.

Elvevag, B., Foltz, P., Weinberger, D. R., & Goldberg, T. E. (2001). Analysis of clinical interviews of patients with schizophrenia. *Eleventh Annual Meeting of the Society for Text and Discourse*. University of California, Santa Barabara.

Blei, D.M., Griffiths, T.L., Jordan, M.I., & Tenenbaum, J.B. (2004). Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems* **16**.

Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, **23**, 229.

Press, W. H., Flannery, B., Teukolsky, S., & Vetterling, W. (1986). *Numerical Recipes.* Cabridge University Press, Cambridge.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240.

Methods

For each corpus, we construct a matrix whose elements, M_{ij}, are given by

$$M_{ij} = \ln(m_{ij} + 1)S_j$$

where m_{ij} is the number of times that the jth word is found in the ith document, and

$$S_j = 1 + \frac{\sum_{i=1}^{N} P_{ij} \ln(P_{ij})}{ln(N)}$$

is the weight given to each word, which depends on the information entropy of the word across documents (Dumais, 1991). In the above expression

$$P_{ij} = \frac{m_{ij}}{\sum_{i=1}^{N} m_{ij}}$$

is the probability density of the jth word in the ith document, and N is the total number of documents in the corpus (Dumais, 1991).

Given the weighted matrix, M, we then construct a reduced representation by performing Singular Value Decomposition (SVD) and keeping only the singular vectors which correspond to the n largest singular values. This step relies on a linear algebra theorem which states that any $M \times N$ matrix A with $M > N$ can be written as $A = USV^T$, where U is an $M \times N$ matrix with orthonormal columns, V^T is an $N \times N$ matrix with orthonormal rows, and S is a $N \times N$ diagonal matrix (Press, Flannery, Teukolsky, & Vetterling, 1986). By writing the matrix equation as

$$A_{ij} = \sum_{l=1}^{N} U_{il} S_l V_{jl}$$

it is clear that for a spectrum of singular values S_l which decays in some well-behaved way, the matrix A can be approximated by the n highest singular values and corresponding singular vectors. In this reduced representation, the dot product of the rows (or columns) of A can be calculated by using the left and right singular vectors scaled by the singular values, e.g. $A^T A = (USV^T)^T(USV^T) = (VS)(VS)^T$, since U and V have orthonormal columns. In typical applications best results are obtained by keeping ∼300 singular values (Landauer, Foltz, & Laham, 1998). The number of singular values that we keep in the five corpora ranges from 300 to 420.

In calculating the correlation dimension of the corpora, we use the normalized, rather than the full, document vectors. The choice is motivated by the observation that the measure of similarity between documents used in Latent Semantic Analysis (LSA) applications is the cosine of the angle between the two vectors. LSA can be very successful in some document comparison tasks, such as assigning grades to essays based on the cosine distance of each essay from various prescored ones (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). By using the cosine of the angle between the two document vectors as the similarity measure, the method de-emphasizes the importance of vector length to the measure of semantic distance. Vector length is associated with the length of the paragraph the vector represents; two paragraphs can be semantically very similar, while being of significantly different length. Geometrically, this is equivalent to considering the semantic distance between the projection of the document vectors onto the unit n-dimensional sphere.

To construct the randomized English corpus, we built each paragraph in turn by taking at random, and without replacement, a word from the corpus until we reach the length of the original paragraph, and we repeat the process for all the paragraphs. It is thus clear that the randomized corpus contains the exact number of paragraphs and words as the original, and that all word frequencies are also exactly the same, however, the word choice for each paragraph has been permuted.